基于序列模式挖掘的入侵检测系统的研究

孟宪苹,宋 菲,李 俊
(南京航空航天大学 信息科学与技术学院,江苏 南京 210016)

摘 要:对入侵检测和数据挖掘从定义和分类等各方面等进行了基本介绍,提出了一个基于数据挖掘的入侵检测系统的总体框架,其整个系统分为训练阶段和测试阶段,对其中各个模块进行基本的功能分析。为了提高数据挖掘的效率,可以将序列模式挖掘引入该入侵检测系统中。将关联规则算法和序列模式挖掘算法同时使用;增加挖掘的粒度。对序列模式挖掘的算法进行了具体分析,并通过具体的实例来说明引入序列模式挖掘能更好地提高数据挖掘的效率。

关键词:入侵检测系统;数据挖掘;序列模式挖掘
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2008)03-0154-03

Research of Intrusion Detection System Based on Sequential Pattern Mining

MENG Xian-ping, SONG Fei, LI Jun
(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: *Introduces the concepts about the definition and the sort of the intrusion detection system and data mining, presents a total intrusion detection frame based on data mining, it includes two phases, training phase and testing phase, and analyses the function of the each model in the frame. In order to advance the efficiency of the data mining, the sequential pattern mining is inducted into this frame. To increase the mining depth, the sequential pattern mining algorithm can be combined with the association rules algorithm. In the end analyses the sequential pattern mining algorithm, and through the example to explain the advancement of the efficiency of the data mining.*

Key words: intrusion detection system; data mining; sequential pattern mining

1 入侵检测与数据挖掘

入侵检测(Intrusion Detection, ID)技术[1]是指对任何破坏计算机系统的安全性、完整性以及机密性的活动的识别。与传统的操作系统加固、身份认证和防火墙隔离等静态安全防护技术不同,入侵检测作为一种积极主动的动态安全防御技术,它提供了对内部攻击、外部攻击和误操作的实时保护,在网络系统受到危害之前拦截相应入侵,是网络防火墙的有力补充。入侵检测技术依据具体的检测方法可分为异常检测和误用检测[2]。

数据挖掘[3](Data Mining, DM)是一种重要的数据分析方法,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,但又是潜在的有用的信息和知识的过程。它是知识发现(Knowledge Discovery of Database, KDD)过程中的一个关键步骤,是当前涉及人工智能和数据库等学科的一个相当活跃的研究领域。它基于人工智能、机器学习和统计等技术,能分析原有的数据,做出归纳性的推理,从中挖掘出潜在的模式,预测出客户的行为。常用于入侵检测的数据挖掘的分析方法主要有关联分析方法、分类分析方法、聚类分析方法和序列模式分析方法。

2 系统模型及功能

2.1 系统模型

入侵检测系统分别工作在训练阶段和测试阶段。其系统模型如图1、图2所示。

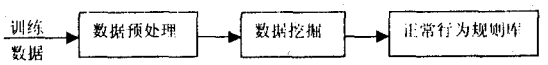

图1 训练阶段

2.2 各模块功能

如图1所示,在训练阶段,输入训练数据,经过数

收稿日期:2007-06-16
作者介绍:孟宪苹(1983-),女,山东滕州人,硕士研究生,研究方向为计算机网络;李 俊,教授,硕士生导师,研究方向为计算机网络和数据库。

据预处理将原始数据处理成适合数据挖掘的数据记录,包括数据清洗等。在当前时间窗口内,用关联规则算法挖掘出频繁规则,对前 K 个时间窗口内的数据记录使用序列模式挖掘算法挖掘网络事件之间的行为规则,从而得到正常行为规则库。

如图2所示,在检测阶段,主要分为以下8个模块:

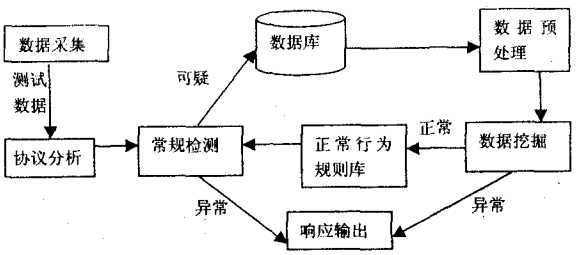

图2　测试阶段

(1)数据采集模块:使用 winpcap 将网卡设置成混杂模式,进行数据采集。

(2)协议分析模块:对数据采集模块捕获的数据包进行协议解析,提取包头中的信息,其中信息包括 Time_ Stamp, Service, Src_ host, Dest_ host, Src_ pot, Dest_ pot, Src_ bytes, Dest_ bytes, Flag 等。

(3)常规检测模块:通过规则库中设置的静态规则可实现一些常规的、简单的网络攻击的检测。

(4)数据库:存储测试数据和实时网络数据的属性和信息。若常规检测未发现异常,则将数据包的重要信息和属性存入数据库中。

(5)数据预处理:数据预处理的主要功能是数据清洗。主要目的是将数据库中的无用数据清理掉,并修正可能出现的数据错误。将原始数据处理成适合数据挖掘的数据记录。

(6)数据挖掘:从数据库中提取数据,然后分别对当前窗口和前 K 个时间窗口内的数据记录调用关联规则算法和序列模式算法,挖掘出可疑行为规则,判断该可疑行为是否是异常行为,如果是异常行为或者概率低于一定阈值则判定为异常行为,响应输出。

(7)正常行为规则库:存储静态规则和由测试数据挖掘出的规则。常规检测时从规则库提取静态规则进行检验。数据挖掘时提取测试数据得到的规则进行检测。

(8)响应输出:对检测结果进行响应,并将响应结果输出。

3　序列模式挖掘

3.1　序列模式挖掘的概念

关联规则挖掘基本概念包括[4]:

(1)一个序列(Sequence)是项集的有序表,记为 $a = a_1 \rightarrow a_2, \cdots, a_n$,其中每个 a_i 是一个项集。一个序列的长度(Length)是它所包含的项集。具有 k 长度的序列称 k - 序列。

(2)设序列 $a = a_1 \rightarrow a_2, \cdots, a_n$,序列 $b = b_1 \rightarrow b_2, \cdots, b_m$,若存在整数 $i_1 < i_2 < \cdots < i_n$,使得 a_1 包含于 b_{i1}, a_2 包含于 $b_{i2}, \cdots, a_n$ 包含于 b_{in},则称序列 a 是序列 b 的子序列,或序列 b 包含序列 a。在一组序列中,如果某序列 a 不包含在其他任何序列中,则称 a 是该组中最长序列(Maximal sequence)。

(3)给定序列 S,序列数据库 D_T,序列 S 的支持度(Support)是指 S 在 D_T 中相对于整个数据库元组而言所包含 S 的元组出现的百分比。支持度大于最小支持度(min - sup)的 k - 序列,称为 D_T 上的频繁 k - 序列。

(4)序列模式也称序列关联,可表示成如下形式:

When A occurs => B occurs within some certain time

(5)频繁有效事件(frequent episode)可表示成如下形式:

$$X, Y \rightarrow Z[c, s, w] \quad (1)$$

其中 X, Y, Z 是项集;c 是可信度,s 是支持度:

$$s = \text{support}(X \cup Y \cup Z) \quad (2)$$

$$c = \text{support}(X \cup Y \cup Z)/\text{support}(X \cup Y) \quad (3)$$

w 为时间间隔,对于给定的一个有时间戳的标记的时间记录集,每个记录是一些项的集合,时距 $[t_1, t_2]$ 表示事件序列从 t_1 开始,t_2 结束,时距的宽度定义为 $w = t_1 - t_2$,则 w 表示规则每次出现都必须在 w 范围内。

(6)局部有效事件规则(serial episode rule),是指 X, Y, Z 在事务发生过程中遵循局部时间顺序,比如说 Z 在 Y 之后,并且 Y 在 X 之后。

3.2　序列模式挖掘算法

目前主要的序列模式挖掘算法包括:AprioriAll, AprioriSome, GSP, SPADE, Prefixspan 等。

AprioriAll 算法的主要框架[4,5]为:

输入:大项集阶段转换后的序列数据库 D_T

输出:所有最长序列

$L1 = \{\text{large } 1 - \text{sequences}\}$;//大项集阶段得到的结果

for ($k = 2; L_{k-1} \neq \emptyset; k++$)do begin

$C_k = \text{AprioriALL} - \text{generate}(L_{k-1})$;//$C_k$ 是从 L_{k-1} 中产生的新的候选者

for each customer - sequence c in D_T do//对于在数据库中的每一个客户序列 c

Sum the count of all candidates in C_k that are

contained in c;//对包含于 c 中 C_k 内的所有候选者记数

L_k = Candidates in C_k with minimum support;//L_k = C_k 中满足最小支持度的候选者

end

Answer = Maximal Sequences in $\bigcup_k L_k$;

上面算法的关键是候选集的产生,具体候选者的产生如下:

AprioriALL - generate( ) 函数 //计算候选者的产生

输入:所有的大(k - 1) 序列的集合 L_{k-1}

输出:候选 C_k

insert into C_k//首先进行 L_{k-1} 与 L_{k-1} 的连接运算

select $p.\text{litemset}_1$, $p.\text{litemset}_2$,…, $p.\text{litemset}_{k-1}$, $q.\text{litemset}_{k-1}$

from $L_{k-1}p$, $L_{k-1}q$//p, q 是 L_{k-1} 中不同的序列串

where $p.\text{litemset}_1 = q.\text{litemset}_1$,…, $p.\text{litemset}_{k-2} = q.\text{litemset}_{k-2}$;//下一步删除 $c \in C_k$ 所有序列,且 c 的某些序列(k - 1) 就不在 L_{k-1} 中

for 所有 $c \in C_k$ 的序列 do

for 所有 c 的(k - 1) 序列 do

if ($s \in L_{k-1}$) then delete 来自于 C_k 的 c

3.3 序列模式挖掘结果示例

如表 1 所示,对其数据库进行序列挖掘。

表 1 用户行为时序数据库

CID	TID	ISet
1	2	*bcd*
2	1	*b*
2	2	*abc*
2	3	*bcd*
3	1	*ab*
3	2	*abc*
3	3	*bcd*

设 min_support=3,序列模式的 L 长度为 3。表 2 为按照上述算法生成的频繁 1 - 序列、2 - 序列、3 - 序列集合。

表 2 各频繁序列集

C_1		L_1		C_2		L_2		C_3		L_3	
1 - 序列	支持度	1 - 序列	支持度	2 - 序列	支持度	2 - 序列	支持度	3 - 序列	支持度	3 - 序列	支持度
a	3	*a*	3	*ab*	3	*ab*	3	*abc*	2	*bcd*	3
b	7	*b*	7	*ac*	2	*bc*	5	*abd*	0		
c	5	*c*	5	*ad*	0	*bd*	3	*acd*	0		
d	3	*d*	3	*bc*	5	*cd*	3	*bcd*	3		
				bd	3						
				cd	3						

序列挖掘的最后结果为一个频繁 3 - 序列 *bcd*。

4 结束语

文中的创新之处就是提出了一个新的基于数据挖掘的入侵检测系统框架。在系统中,将数据挖掘中的序列模式挖掘应用到入侵检测系统中来,对其中数据挖掘的部分采用关联规则算法和序列模式挖掘算法相结合的方法。

参考文献:

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

[2] 宋世杰.基于序列模式挖掘的误用入侵检测系统框架研究[J].计算机工程与科学,2006(1):28 - 30.

[3] Bace R G. Intrusion Detection[M]. US:Macmillan Technical Publishing,1999.

[4] 李川川,刘衍珩,田大新.基于序列模式挖掘的网络入侵检测系统[J].吉林大学学报,2007(1):121 - 125.

[5] 钱 昱,郑 诚.基于序列模式的异常检测[J].微机发展,2004,14(9):53 - 55.

(上接第 141 页)

[2] Cortes C, Vapnik V. Support - vector networks[J]. Machine Learning,1995,20(3):273 - 297.

[3] Pdesa M J.模式识别——原理、方法及应用[M].吴逸飞译.北京:清华大学出版社,2002.

[4] 刘志刚,李德仁,秦前清,等.支持向量机在多类分类问题中的推广[J].计算机工程与应用,2004(7):10 - 13.

[5] Arenas - Gsrcia J, Perez - Cruz F. Multi - class support vector machines:A new approach[C]//ICASSP,2003. Hong Kong:[s.n.],2003.

[6] Xu P, Chan A K. Support vector machines for multi - class signal classification with unbalanced samples[C]//Proceedings of the International Joint Conference on Neural Networks 2003. Portland:IEEE,2003:116 - 119.

[7] Bottou L, Cortes C, Denker J, et al. Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition [C]//Proc of the Int Conf on Pattern Recognition. Jerusalem:[s.n.],1994:77 - 87.

[8] Krebel U. Pairwise Classification and Support Vector Machines [C]//In:Scholkopf B, Burges C J C, Smola A J, eds. Advances in Kernel Methods: Support Vector Learning. Cambridge, MA:The MIT Press,1999:255 - 268.

[9] 张爱丽,刘广利,刘长宇.基于 SVM 的多类文本分类研究[J].情报学报,2004(9):6 - 10.