基于规则约束制导的入侵检测研究

李 雷,丁亚丽,罗红旗
(南京邮电大学 自动化学院,江苏 南京 210003)

摘 要:入侵特征值识别和发现算法是误用入侵检测中的关键技术。采用数据挖掘技术从主机和网络的数据中发现入侵特征值,建立入侵行为和正常行为规则库,已经广泛用于入侵检测技术中。针对数据挖掘中经典的Apriori和AprioriTid算法中存在项集生成瓶颈问题,提出了一种基于规则约束制导的Apriori算法,考虑到强规则事件并不一定是有趣事件并且大部分入侵行为都是基于已有入侵模式基础上变异得到,加入兴趣度约束和递减支持度约束。通过实验演示,结果表明该算法可大幅提高效率并在入侵检测漏报率上有很好的改善。

关键词:数据挖掘;入侵检测;递减支持度;关联规则;频繁项集
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2010)03-0143-04

Intrusion Detection Technology Research Based on Homing-Constraint Rule

LI Lei,DING Ya-li,LUO Hong-qi
(College of Automation, Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:Network security has been a very important issue, since the rising evolution of the Internet. One commonly used defense measure against such malicious attacks in the Internet is Intrusion Detection System (IDS). Data mining has been extensively applied in network intrusion detection and prevention systems by discovering user behavior patterns from the network traffic data. Association rules and sequence rules are the main technique for data mining. Considering the classical Apriori algorithm and AprioriTid algorithm with two bottlenecks of frequent itemsets mining, this paper proposed a homing-constraint-rule Apriori algorithm (HCRAporiori). Experiment results indicate that the proposed method is efficient.

Key words:data mining;intrusion detection; descending support;association rules; frequent itemsets

0 引 言

传统的网络安全技术,如防火墙、加密技术,实现的是"分而治之"解决方法,是网络安全防护系统构成的环节。从实现的防护能力上讲,这些技术实现的是一种静态的被动防护,其安全防护的层次在网络的边界,能阻止大部分的外部攻击,但是对内部攻击却无能无力。网络入侵检测[1]是从计算机网络系统的若干关键点收集信息,并分析网络中是否存在入侵行为和迹象。网络入侵检测技术主要有两种基本类型,即异常检测和误用检测。异常检测的主要优点是不需要过多的有关系统缺陷的知识,并且能够检测出未知的入侵模式。

近年来,针对异常检测有了较多研究。文献[2]提出基于隐Markov模型(HMM)的用户行为异常检测方法,这种方法的检测准确率较高,但检测效率较差。文献[3]提出基于HMM的程序行为异常检测方法,该方法在训练数据充足的情况下能够获得比较高的检测准确率,但计算成本很高。

采用数据挖掘技术从主机和网络的数据中发现知识,建立入侵行为和正常行为规则库,并在实时检测中利用数据挖掘抽取用户和系统行为模式[4],以进行检测。系统能有效识别并自动更新规则库,提高检测系统的可扩展性和自适应性,有效降低误报率和漏报率。

1 研究背景

入侵检测根据所采用的技术分为误用检测和异常检测。其核心都是分析入侵行为或正常行为的特征,建立模式库。模式库是IDS[5]中的关键部件,模式的

收稿日期:2009-07-02;**修回日期**:2009-10-26
基金项目:国家自然科学基金项目(10371106,10471114);江苏省高校自然科学基金项目(04KJB110097,08KJB520023);南京邮电大学攀登计划项目(NY207064)
作者简介:李 雷(1958-),男,安徽砀山人,教授,研究方向为智能信号处理、非线性分析与计算智能。

准确性及完备性将直接影响到入侵检测的性能。

关联规则挖掘是一种有效的数据挖掘方法，它能够提取数据库特征项之间的相互关系，将其应用于入侵检测，可快速、高效地建立模式库。挖掘关联规则的问题即寻找一个支持度和置信度均大于用户指定的最小支持度 S 和最小置信度 C 的规则。一般分为两步：

①找出所有支持度大于用户最小支持度的频繁项集；

②根据所得的频繁项集生成需要的关联规则。

其中挖掘频繁项集是关键问题，以 R. Agrawal 等人提出的 Apriori[6]算法和 AprioriTid 算法[7,8]为代表。

2 规则约束制导算法

数据挖掘[9]过程可以从给定的数据集中发现数以千计的规则，其中大部分规则与用户不相关或用户不感兴趣。通常，用户具有很好的判断能力，知道沿什么“方向”挖掘可能得到有趣的模式，从而知道他们感兴趣的“形式”模式或规则。这样，一种好的启发式方法是让用户利用他们的直觉或期望作为限制搜索空间的约束条件。这种策略称作为基于约束的挖掘。对于入侵检测系统而言，在如下几个方面加以考虑，从而提出规则约束制导算法，并将其应用于入侵检测的模式建库行为中。

首先，考虑到现实生活中，许多入侵行为或病毒均基于原先的模式进行变异。典型的如：Code Red I 和 Code Red II，它们都属于蠕虫病毒。Code Red 病毒的主要攻击对象为 Microsoft 的 IIS Web Server，更精确地说，是 IIS Web Server 的一只 ISAPI Extension 程式。Code Red I 向网站发送大量消息以至最后不得不关闭该网站，这就是所谓的拒绝服务式攻击(DDoS)。Code Red II 则更加强大且更具破坏性。在 SNORT 中，检测 Code Red I 和 Code Red II 攻击的规则分别为：

(1)检测 Code Red I 的规则。

alert tcp any any - > any 80 (msg:“Code Red I Overflow”; content “|2F646566 1756C74 2E696461 3F4E4E4E|”;)

(2)检测 Code Red II 的规则。

alert tcp any any - > any 80 (msg:“Code Red II Overflow”; content: “|2F646566 1756C74 2E696461 3F585858|”;)

比较两个规则发现 Code Red I 和 Code Red II 含有一个相同特征子“2F646566 1756C74 2E696461”。

由上可知，在实际攻击中，许多攻击特征具有相同特征子串，因此可借助于已知特征，找出新型攻击特征。其次，大部分基于关联规则的挖掘算法都使用支持度 - 置信度框架。尽管最小支持度和置信度阈值有助于排除大量无趣规则的探查，但是仍然会产生一些用户不感兴趣的规则。下面来看一个强关联规则但没有趣的例子。

假定对分析涉及购买计算机游戏和录像的 All Electronics 事务感兴趣。设 game 表示包含计算机游戏的事务，而 video 表示包含录像的事务。在所分析的 10000 个事务中，数据显示 6000 个顾客事务包含计算机游戏，7500 个事务包含录像，而 4000 个事务同时包含计算机游戏和录像。假定发现关联规则的数据挖掘程序对该数据运行，使用最小支持度 30%，最小置信度 60%，发现下面的关联规则：

$$\text{buys}(X,\text{“computergames”}) \Rightarrow \text{buys}(X,\text{“video”}) \quad [\text{support}=40\%, \text{confidence}=66\%] \quad (1)$$

规则(1)是强规则，因而提出报告，因为其支持度的值为 4000/10000 = 40%，置信度的值为 4000/6000 = 66%，分别满足最小支持度和最小置信度阈值。然而，规则(1)是误导，引文购买录像的概率是 75%，比 66% 还大。事实上，计算机游戏和录像时负相关的，因为买一种实际上减少了买另一种的可能性。不完全理解这种现象的话，容易根据规则作出不明智的商务决定。因此不一定是强规则事件就是有趣的。

最后，基于所有入侵行为都据有时间序列性，它是由一系列行为操作所组成的，所以在挖掘频繁项时，应该考虑其时间属性。比如{A,B,C}组成的一事务，B 是在 A 后面发生，同理，{A,B}决定 C，而不是{A,C}后发生 B。

基于以上三点，提出一种规则约束制导算法，将经典的 Apriori 算法加入已知特征序列来检测其衍生的入侵方式。考虑到第二点，在较低层使用递减最小支持度，同时加入时间序列性，提出基于规则约束制导的 HCRAporiori 算法。

3 HCRApriori 的算法流程

HCRApriori 算法包含的属性集：D 为事务数据集合，记为 tid，C 为候选项集，$S(c)$ 为候选项集 C 的支持度，ID 为项目集合记为 id，time 是各项集之间的时间属性，des_ sup 为满足阈值的递减支持度，known_ signature 为已知攻击特征，max_ len 为包含已知特征的频繁项集的最大长度，L_k 为具有满足该层最小支持度的频繁项集 k_ itemsets 集合，sig 为集合 L_k 中的当前频繁项集，C_k 为候选频繁项集 k_ itemsets 的集合。

算法 HCRApriori 由下面几个方面组成。

第一块：具有规则约束制导的 Apriori 算法。

输入：D 数据事务库

des_ sup 递减支持度 max_ des_ sup - -

输出:L_k 满足要求的频繁项集 k - 项集

方法:

(1)L_1 = find_ frequent_ 1 - itemsets(D,max_ des_ sup)

(2)L_k = find_ frequent_ k - itemsets(D,max_ des_ sup - -,known_ signature,time)

(3)for k = n + 1 to max_ len do

(4)C_k = candidates_ gen(D,$L_1 \infty L_{k-1}$,max_ des_ sup - -)

第二块:find_ frequent_ 1 - itemsets 算法。

输入:D 数据事务库

max_ des_ sup

输出:$L_1$1 - 项集

方法:

(1)$L_1 = \varnothing$

(2)while (D 存在事务) do begin

(3) 从 D 中读出一个事务 tid

(4){对事务 tid 中每一个事务项 id 分别对应计数,同一个事务 tid 中的事务项只计数一次}

(5)end

(6)$L_1 = \{c \in C_1 \ \& \ S(c) \geq$ max_ des_ sup$\}$

第三块:find_ frequent_ k - itemsets 算法。

输入:D 数据事务库

des_ sup 递减支持度 max_ des_ sup - -

$L_1$1 - 项集

known_ signature 已知特征

输出 L_kk - 项集

方法:

(1)$L_k = \varnothing$

(2)for(L_1 中的每个项 id) do

(3){添加 known_ signature 的 id 到候选项集合 C_k}

(4)$L_k = \{c \in C_k \ \& \ S(c) \geq$ max_ des_ sup - -$\}$

第四块:candidate_ gen 算法。

输入:D 数据事务库

des_ sup 递减支持度 max_ des_ sup -

$L_1$1 - 项集

方法:

(1)$L_k = \varnothing$

(2)for(L_1 中的每个项 id) do begin

(3){添加(sig + id) 到候选项集合 C_k

(4)$L_k = \{c \in C_k \ \& \ S(c) \geq$ max_ des_ sup - -$\}$

(5)if 不存在任何一个 S(sig + id) $\geq$ max_ des_ sup - - then do

(6)end}

(7)return L_k

4　实验结果

实验环境:Intel Core(TM)2 Duo Processor 1.83GHz with 1024 RAM;操作系统为 Window XP。算法用 C + + 实现。

实验数据:DARPA1999[10]数据。在我们的实验中,前两周数据作为训练数据,其中第一周数据不含有任何攻击行为,第二周只包含 43 个攻击实例。通过已知特征,那些攻击行为大致可以分为 18 种。

实验一:设定 known - signature 长度为 2,数据库 D 大小为 5M,比较经典 Apriori 算法和 HCRApriori 算法在生成 n - itemsets 运行时间,经过重复多次试验后得出图 1。

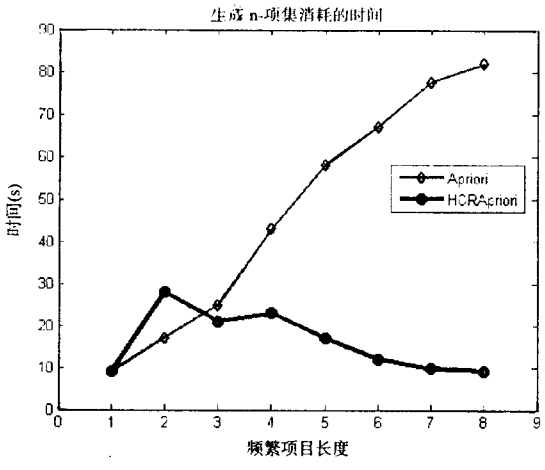

图 1　不同的 n - 项集消耗时间对比

实验分析:从图 1 曲线可知,随着 n - 项集的 n 值逐渐变大,HCRApriori 算法消耗的时间远远小于经典 Apriori 算法,是因为经典 Apriori 算法消耗了不少时间在生成大量的频繁项集合,而这些集合有绝大部分不是入侵行为模式的频繁项结合,是无趣事件。在生成候选项集 C_k 时,采取用 $C_k = L_{k-1} \infty L_1$ 得到,减少浏览数据库的次数和减小候选项集合的大小,从而也节约了不少时间。同时发现在 2 - 项集处,HCRApriori 算法消耗的时间比 Apriori 算法时间多,那是因为在 n = 2 处,HCRApriori 算法多个模式匹配的过程,同时递减支持度导致候选集合中满足支持度的项集略为膨胀。综观全局 HCRApriori 算法无论在时间消耗还是在生成有趣事件上都比 Apriori 算法有相当的优越性。

实验二:比较 Apriori 算法,无递减支持度 HCRApriori 算法记为(NApriori)[5]和 HCRApriori 算法在误报率的优越性。数据库 D = 10M,HCRApriori 递

减规则为先设定 max_des_support 为上层最小支持度,其次为每隔两层取上层最小计数为次层的最小支持度。已知特征长度取 4。实验结果如图 2 所示。

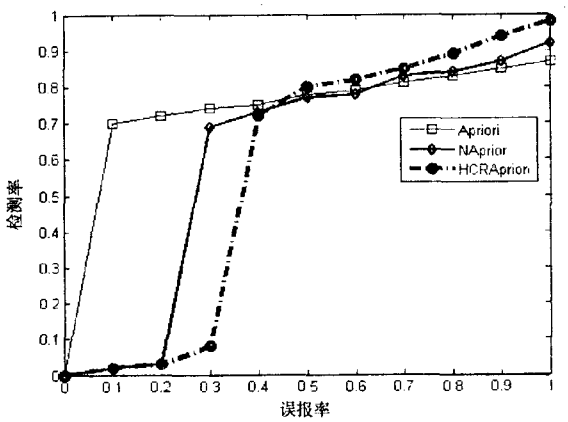

图 2 误报率检测率性能对照图

实验分析:结果显示,使用具有递减支持的 HCRApriori 算法比 Apriori 算法 和 NApriori 算法具有较高精确的检测性能。同样的检测率下,HCRApriori 算法比其他两个算法具有更低的误报率。在试验中,也发现使用较短特征曲线的检测率比较长特征曲线的检测率要高。原因是正常流量中的较短特征出现的概率偏高,但会导致较高的误报率。

5 结束语

文中提出了一种基于规则约束制导的数据挖掘算法,利用已知特征来检测现有入侵模式的变异事件,在此基础上加入递减支持度,以免忽略低支持度事件有可能也是兴趣事件。本算法在消耗时间上大幅度降低和在误报率检测率上都有很好的改进。递减支持度规则的优越性,直接影响到检测率的高低。如何选取更优越的递减支持度规则是笔者下一阶段的研究方向。

参考文献:

[1] 阮耀平,易江波. 计算机系统入侵检侧模型与方法[J]. 计算机工程,2005,28(11):232-236.

[2] 高海华,王行愚,杨辉华. 基于群智能和 SVM 的网络入侵特征选择和检测[C]//2005 年中国智能自动化会议论文集[C]. 北京:国防工业出版社,2005:111-114.

[3] 彭竹苗,张正道,白瑞林,等. 基于 HMM 模型的网络入侵误用检测方法[C]//2007 中国控制与决策学术年会论文集. 沈阳:东北大学出版社,2007:67-70.

[4] 王玉震,李 雷. 基于 SVR 的图像增强方法[J]. 计算机技术与发展,2009,19(1):60-62.

[5] Hu Zhengbing, Li Zhitang, Wu Junqi. A novel network intrusion detection (NIDS) based on signatures search of data mining[C]//2008 Workshop on Knowledge Discovery and Data Mining. Moscow, Russia:[s.n.],2008:10-16.

[6] Agrawal R, Imieilnski T, Swami A. Mining association rules between sets of items in large databases[C]//Proc of the ACM SICMOD Int Conf on Management of Data. Washington DC:[s.n.],1993:207-216.

[7] Han J W, Pei J, Yin Y. Mining frequent paterns without candidate generation[C] //Proc of the 2000 ACM SIGMOD Int Conf on Management of Data. Dallas:[s.n.],2000:1-12.

[8] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc of the 20th Into Conf Very Large Data Bases. Santiago, Chile:[s.n.],1994:487-499.

[9] 韩家炜,坎 伯. 数据挖掘:概念与技术[M]. 范 明,孟小峰译. 北京:机械工业出版社,2001:152-157.

[10] KDD (1999), the third international knowledge discovery and data mining tools competition data set (KDD99 Cup)[EB/OL]. 1999. http://kdd. ics. uci. edu/databases/kddcup99. html.

(上接第 142 页)

(2)利用人眼掩蔽特性根据小波低频系数块所属类别选取相应的量化步长,并且水印嵌入过程中块内系数的改变量自适应于系数本身大小,具有较好的自适应性。

(3)对各种常见攻击具有良好的鲁棒性,并且水印的提取不需要原始图像,具有较好的实用性。

参考文献:

[1] 董 敏,王向阳. 基于分块量化的小波域数字水印嵌入算法[J]. 微电子学与计算机,2007,24(7):31-34.

[2] Masry M, Ramos M, Hemami S. Robust data hiding using psychovisual thresholding[C]//ICIP2000 Proc 2000 Int Conf Image Processing. Vancouver, Canada: IEEE Signal Processing Society,2000:593-596.

[3] Cox I J, Killian J. Secure spread spectrum watermarking for multimedia[J]. IEEE Transactions on Image Processing, 1997,6(12):1673-1690.

[4] 李赵红,侯建军. 基于 Logistic 混沌映射的 DCT 域脆弱数字水印算法[J]. 电子学报,2006,34(12):2134-2137.

[5] Sepsirisuk K. An adaptive digital watermarking based on a tree structure using the human visual system[J]. IEEE IN T,2005 (2):1062-1065.

[6] Nillnb A. Visual model weighted cosine transform for image compression and quality assessment[J]. IEEE Transaction on Communications,1995,33(6):551-557.

[7] 朱兴力,张家树. 基于小波系数块能量分析的自适应数字水印算法[J]. 计算机应用,2006,26(4):830-832.

[8] Liu Hongmei, Huang Jiwu. An adaptive video watermarking algorithm in wavelet domain[J]. Acta Electronica Sinica, 2001,29(12):1656-1660.