一种面向入侵检测的数据挖掘算法研究

叶和平,尚 敏
(广东科学技术职业学院 软件工程系,广东 广州 510640)

摘 要:为提高入侵检测的精确性和有效性,通过对基本序列模式挖掘算法(Aprior算法)的分析,针对其缺点并结合入侵检测数据的特殊性,设计了改进的Aprior算法用于序列模式挖掘算法,算法将数据属性分成多个等级,侧重于多属性的序列模式挖掘,算法首先寻找高频轴属性值事件,再迭代降低支持度并增加新的低频轴属性值,用于比较长的频繁项集。同时以网络数据和日志文件数据为实验基础,从算法的精确性和适应性方面进行了比较。

关键词:Aprior算法;序列模式挖掘;入侵检测;数据挖掘
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2008)11-0149-03

Study on an Intrusion Detection Oriented Data Mining Algorithm

YE He-ping, SHANG Min
(Dept. of Software Eng., Guangdong Vocational Institute of Science and Technology, Guangzhou 510640, China)

Abstract: An improved sequential patterns mining algorithm based on Aprior algorithm for intrusion detection is designed. It classifies data by the properties and focus on the sequential patterns by the multiple properties. It is effective for the long frequent item sets and improves the accuracy of intrusion detection largely. Its properties of accuracy and adaptation have been verified by analysis of audit data from network record and log files.

Key words: Aprior algorithm; sequence patterns mining; intrusion detection; data mining

0 引 言

计算机网络在日常生活中扮演着越来越重要的角色,与此同时,出于各种目的,它正日益成为犯罪分子的攻击目标,黑客们试图使用他们所能找到的方法侵入他人的系统。为此,必须采取有效的对策以阻止这类犯罪发生。开发具有严格审计机制的安全操作系统是一种可行方案,然而综合考虑其实现代价,在许多问题上作出少许让步以换取减少系统实现的难度却又是必要的。因此,在操作系统之上,再加一层专门用于安全防范的应用系统成为人们追求的目标。入侵检测技术即是这样一种和其它安全技术一道构成计算机系统安全防线的重要组成部分之一。自从Dorothy E. Denning1987年提出入侵检测的理论模型后[1],关于入侵检测的研究方法就层出不穷,特别是对异常检测的研究,借助于相关学科理论的进展,人们提出了许多解决办法[2~8]。文中通过引入数据挖掘技术中的序列模式挖掘方法,设计了一种改进的基于Aprior算法的序列模式挖掘算法,并从算法的精确性、适应性对算法的优劣进行评价。

收稿日期:2008-02-31
基金项目:广东省自然科学基金(04010589)
作者简介:叶和平(1968-),男,讲师,研究方向为信息安全,网络与并行分布式算法;尚 敏,教授,研究方向为计算机应用。

1 序列模式挖掘算法

序列分析算法的思想是,获取数据库记录之间在时间窗口中的关系。这类算法可以发现审计数据中的一些经常以某种规律出现的事件序列模式。这些频繁发生的事件序列模式可帮助在构造入侵检测模型时选择有效的统计特征[9]。

测的基本过程为信息收集、信息分析和结果处理。

1.1 序列模式挖掘算法SPM-ID

序列模式挖掘时,数据源由基于频繁网络模式和频繁系统活动模式[10,11]的单个审计数据流中获得,因而传统从事件流数据中获取单序列模式的算法[12]及不同的多数据序列中获取多个序列模式的算法都不再适用。下面对入侵检测中序列模式挖掘模型[12]基于轴属性、参考属性、相关支持度的序列模式挖掘算法SPM-ID(Sequential Patterns Mining for Intrusion Detection)进行分析。这里先给出数据挖掘的相关理论。

定义1 非空集合 $I=\{i_1,i_2,i_3,\cdots,i_m\}$ 称为项集,其中 i_k 称为项。

定义 2 序列是项集的有序表，记为 $a = a_1 \rightarrow a_2 \rightarrow \cdots \rightarrow a_n$，其中 $a_k \subset I\ (k = 1, \cdots, n)$，含有 k 个项的序列长度为 k，称为 k 序列($k = \sum |a_i|$)。

定义 3 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

定义 4 判断序列模式是否有效的参数称为模式浓度(Pattern's strength)。

模式浓度通常基于在给定数据下模式发生的频率来计算。如果模式出现得足够频繁，就称它为有效模式。决定模式是否有效的参数有：① 支持度；② 可信度；③ 重要性；④ 覆盖度。对于给定的一个有时戳标记的事件记录集，每个记录是一些项的集合，时距$[t_1, t_2]$事件序列从 t_1 开始，t_2 结束，时距的宽度定义为 $w = t_2 - t_1$。以 X 表示是项集，那么一个时距就表示包含 X 的一次最小出现，也就是说该时距的任何子时距都不包含 X 的出现。

定义 5 频繁有效事件(frequent episode) 可表示为如下形式：

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

其中，X, Y, Z 是项集；s 是支持度，$s = \text{support}(X \cup Y \cup Z)$；$c$ 是可信度，$c = \dfrac{\text{support}(X \cup Y \cup Z)}{\text{support}(X \cup Y)}$；$w$ 表示规则每次出现都必须在 w 范围内。

定义 6 顺序有效事件规则(serial episode rule) 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。

在面向 IDS 的序列模式挖掘中，从网络数据包和系统日志及审计数据中获取信息时，数据量非常庞大，并且存在杂乱性、重复性和不完整性等问题，因而有必要对数据进行预处理[10]。从网络流上截获的数据主要包括网段内部主机之间，以及网络内部主机和外部主机之间通信的数据包头及数据内容。对这类数据的预处理主要集中在提取几个方面的特征值：网络连接特性、连接的内容特性、连接的统计特征。

对系统日志和审计数据预处理的方式有：数据集成、数据清理、数据简化。

1.2 序列分析

IDS 中进行序列模式挖掘时由于频繁网络模式和频繁系统活动模式只能在网络或操作系统的单个审计数据流中获得，因而传统从事件流数据中获取单序列模式的算法，以及从不同多数据序列中获取多个序列模式的算法都不再适用。针对 IDS 中数据的特殊性，设计了一种新的序列模式挖掘算法，称之为 SPM—ID 算法。

SPM—ID 算法的基本思想是：由于频项集的子集也是频繁的，可以充分利用关联规则挖掘算法中的数据结构和库函数，来挖掘长度≥2 的频繁有效事件；关联规则算法中的原始矢量(raw vector)被用作时间间隔矢量，它的每对值是时间间隔的临界值；最小、无过载的时间连接函数(temporal Join)用于从两个长度为 $k-1$ 的频繁项集时间间隔矢量来创建长度为 k 的侯选项集时间间隔矢量。计算频繁序列模式算法可分为两个步骤：

①通过“轴”特征找到频繁关联；

②在已有的频繁关联的基础上生成频繁序列模式。

1.2.1 基于属性的有效性测量

在不考虑任何专业背景的前提下，传统 Apriori 算法默认用最小支持度和最小可信度来判断模式是否有效。也就是说，假定 I 是模式 p 的有效性测量，那么

$$I(p) = f(\text{support}(p), \text{confidence}(p))$$

其中 f 是一些分等级函数。

如果要将计划级信息加入到有效性测量中，假设 IA 是包含特殊属性的模式 p 的有效性测量，那么

$$Ie(p) = fe(IA(p), f(\text{support}(p), \text{confidence}(p))) = fe(IA(p), I(p))$$

其中 fe 是考虑了模式中属性的分等级函数。审计数据计划级特征的表现形式是“哪些属性必须考虑”，也就是说找到哪些属性可作为引导来挖掘相关的特征。

1.2.2 基于轴属性的分析

在审计数据的属性中存在重要性的排序。对于描述数据而言，有些属性是必需的，有些属性则是辅助性质的。

$Ie(p) = 1$(如果 p 包含轴属性) 或 $Ie(p) = 0$(如果 p 不包含轴属性)

在实际应用中，并不是所有的必要属性都是轴属性。有些网络分析任务需要不同网络服务的静态信息，而其他一些任务需要和主机相关的模式。这些时候，就要选择适当的轴属性。对于频繁有效时间来说，用轴属性限制项的生成是至关重要的。为此引入如下定理。

定理 1 s 是关联规则 $A \rightarrow B$ 的支持度，N 是有效事件规则的总数，这些有效事件规则有如下表现形式：

$$(A \mid B)(, A \mid B)^* \rightarrow (A \mid B)(, A \mid B)^*$$

为避免产生大量无用的事件规则，基本的频繁事件规则算法扩展为频繁序列模式挖掘算法。算法分为两步：

①通过“轴”特征找到频繁关联；

②在已有的频繁关联的基础上生成频繁序列模式。

也就是对第二步而言,事件项集(episode itemsets)建立之后,它的项是有关轴属性的关联,并且属性是有值的。

1.2.3 使用参考属性

除"轴"属性外,另外一个笔者感兴趣的系统审计数据的特征是:有些属性是另外一些属性的参考,这些"参考属性"经常携带有关某"主题"(subject)的信息,而其他属性描述同一主题的"行为"(action)。

当寻找序列"行为"模式时,用"主题"作为参考,这种类型的序列模式可表示成如下形式:

(subject = X,action = a),(subject = X,action = b) →

(subject = X,action = c), [confidence,support, window]

基本频繁有效事件算法可扩展到考虑参考属性。即 $Ie(p) = 1$(如果 p 包含参考属性)或 $Ie(p) = 0$(如果 p 不包含参考属性)。

通过如下算法从审计数据中挖掘出序列模式:

Input:database D,最终最小支持度 st,初始最小支持度 si,轴属性(s)

Output: 频繁有效事件规则(frequent episode rules)Rules

Begin

(1)Rrestricted = $\emptyset$;

(2) 扫描数据库 D 形成 L = {frequent l—itemsets that meet st};

(3)$s = si$;

(4)while($s \geqslant st$) do begin

(5) 从 L 中计算有效事件:每个有效事件必须至少包含一个轴属性的值不在 Rrestricted;

(6) 添加新的轴属性的值到 Rrestricted 中;

(7) 添加新的有效事件规则到规则集 Rules 中;

(8)$S = s/2$;

end

end

2 算法应用及分析

由于 Apriori 算法自身特点,对支持度的大小和频繁项的长度变化十分敏感。较小的支持度使更多项目满足用户要求,频繁集元素个数增加,既而增加了候选集的个数,每趟迭代的计算量随之增加,将大大影响算法的性能。Aprior 算法是基于"自底向上"搜索模式、对于可能包含较长项目集的交易数据库,算法需要数量惊人的计算开销,其成指数增长的复杂度基本上决定了 Aprior 算法只能应用于发现相对较短的频繁项目集。这里列举了 SPM－ID 与 Apriori 算法的比较结果。

表 1 实验数据表示在可信度为 2%、滑动窗口为 2 秒的情况下,随着支持度的不同,算法有不同的精确性。

表 1　算法比较

支持度	3%	2%	1.5%	1%
Aprior 算法	45%	52%	77%	90%
SPM－ID 算法	51%	63%	88%	95%

在 UNIX4.0,ReadHat7.2,Win2000Server 下的实验果如表 2 所示,结果数据表明算法在不同的环境下有较好适应性。

表 2　算法可适应性比较

环境	Unix	Linux	Win2000
Aprior 算法	155s	189s	190s
SPM－ID 算法	133s	136s	135s

3 结束语

通过对序列模式挖掘算法的分析,以 Apriori 算法为基础,设计了一类针对系统审计数据的序列模式挖掘算法,该算法克服了传统 Aprior 算法的局限,将数据属性分成多个等级,侧重于基于属性的序列模式挖掘,算法可应用于比较长的频繁项集,大大提高了入侵检测的精确性。最后以网络数据和日志文件为实验基础,从算法精确性、扩展性和适应性方面对相关的检测算法进行了分析比较,但在关联规则的设计上,如何针对系统审计数据,结合相关的知识模式,设计适当的过滤规则,使序列模式挖掘中导出的关联规则更加有效将是进一步研究的内容。

参考文献:

[1] Denning D E. An intrusion detection model[J]. IEEE Trans on Software Engineering,1987,13(2):222－232.

[2] Lane T,Brodley C E. An Application of Machine Learning to Anomaly Detection[R]. USA:Purdue University,1997.

[3] Hoang Xuan Dau,Hu Jiankun,Bertok P. A Multi－layer Model for Anomaly Intrusion Detection Using Program Sequences of System Calls[R]. Australia:RMIT University,2001.

[4] Balajinath B, Raghavan S V. Intrusion Detection through learning behavior model[J]. Computer communications,2001, 24(8):1202－1212.

[5] Lee W,Stolfo S J. Data Mining Approaches for Intrusion Detection[R]. USA:Columbia University,1999.

(下转第 155 页)

较强的抵抗图像缩放攻击和一定的剪切攻击等几何攻击的能力，尤其是抗 JPEG 攻击的能力很强，能够完全抵御 JPEG 攻击，具有较好的实用性。

表 1　对嵌入水印后的图像进行几种常见的攻击后的检测结果

攻击方法		PSNR(dB)	NC
JPEG 压缩	保持品质因子 90	39.76	1
	保持品质因子 60	34.52	1
	保持品质因子 50	32.60	0.99
	保持品质因子 30	29.54	0.93
	保持品质因子 10	24.68	0.85
加噪声	强度为 0.01 的高斯噪声	23.38	0.98
	强度为 0.01 的乘性噪声	30.17	1
	强度为 0.01 的椒盐噪声	28.71	1
	强度为 0.05 的椒盐噪声	20.88	0.83
平滑滤波	均值滤波	25.7662	1
	中值滤波	30.65	1
	高斯低通滤波	35.89	1
	维纳滤波	34.08	1
图像缩放	缩小 4 倍	30.54	1
马赛克效果	2×2 块	29.16	1
	3×3 块	26.77	0.94
锐化	锐化	31.25	1
剪切	左上剪去 1/4	14.27	0.86

参考文献：

[1] Lee Chang - hsing, Lee Yeuan - Kuen. An adaptive digital image watermarking technique for copyright protection[J]. IEEE Trans on Consumer Electronics, 1999, 45(4): 1005 - 1015.

[2] Podilchuk C I, Zeng W. Image - adaptive watermarking using visual models[J]. IEEE Journal on Selected Areas in Communications, 1998, 16(4): 525 - 539.

[3] Shapiro J M. Embedded image coding using zerotrees of wavelet coefficients[J]. IEEE Trans Singal Processing, 1993 (41): 3445 - 3462.

[4] Ntalian K S, Doulam A D, Doulam N D. An automatic scheme for stereoscopic video object based watermarking using qualified significant wavelet trees[J]. Image Processing, 2002(3): 501 - 504.

[5] Innous H, Miyazaki A, Yamamoto A, et al. A digital watermark based on the wavelet transform and its robustness on image compression[C]// Proceedings of the IEEE International Conference on Image Processing, ICIP'98. Chicago, USA: [s. n.], 1998.

[6] Cox I, Kilian J, Leighton T, et al. Secure spread spectrum watermarking for multimedia[J]. IEEE Transactions on Image Processing, 1997, 6(12): 1673 - 1687.

[7] 孙圣和，陆哲明，牛夏牧，等. 数字水印技术及应用[M]. 北京：科学出版社，2004.

(上接第 148 页)

快速提取、模式算法改进，力争开发出基于协议分析的通用入侵检测系统。

参考文献：

[1] 杜建国. 协议分析和命令解析在入侵检测中的应用[J]. 计算机工程与应用，2004(18): 2 - 3.

[2] Postel J. Internet Protocol DARPA Internet Program Protocol Specification [S/OL]. 1981. http://www.ietf.org/rfc/rfc0791.txt? number=791.

[3] Postel J. Transmission Control Protocol DARPA Internet Program Protocol Specification[S/OL]. 1981. http://www.ietf.org/rfc/rfc0793.txt? number=793.

[4] 刘文涛. 网络安全开发包详解[M]. 北京：电子工业出版社. 2005.

[5] 田　伟. 基于协议分析的网络入侵检测系统研究[D]. 南京：南京信息工程大学，2007.

(上接第 151 页)

[6] Lee. A Data Mining Framework for Constructing Features and Models for Intrusion Detection System[D]. USA: Columbia University, 1999.

[7] Kumar S, Spafford E H. An Application of Patter Matching in Intrusion Detection[R]. USA: Department of Computer Science, Purdue University, 1994.

[8] Doak J. Intrusion Detection: The Application of a Feature Selection - A Comparison of Algorithms and the Application of Wide Area Network Analyzer[R]. USA: Department of Computer Science, University of California, 1992.

[9] Lee W, Stroifo S J. Data mining approaches for intrusion detection[C]//Proc of the 7th USENIX Security Symposium. San Antonio, TX: [s. n.], 1998.

[10] Joshi M, Karypis G. A Universal Formulation of Sequential Patterns[R]. USA: Department of Computer Science, University of Minnesota, 1999.

[11] Bace R G. Intrusion Detection[M]. USA: Macmillan Technical Publishing, 1999.

[12] Mannila H, Tovonen H. Discovering generalized episode using minimal occurrences[C]//Proc. of the 2nd Intl. Conf. on Knowledge Discovery in Database and Data Mining. Portland, Oregon: [s. n.], 1996.