改进 Apriori 算法在入侵检测系统中的应用

Improved Apriori Algorithm in Intrusion Detection System Application

欧阳峥峥　丰洪才　（武汉工业学院 湖北武汉 430023）

摘　要： 提出了一种建立入侵检测系统的方法，该方法基于数据挖掘技术，建成后的系统具有可扩展性、自适应性和准确性特点。结合一些网络攻击行为对关联挖掘算法进行了分析，找出符合条件的项集并建立规则库，从而提高入侵检测技术的检测能力。

关键词： 入侵检测　数据挖掘　Apriori

1　引言

入侵检测系统是安全技术的核心，是防火墙的重要补充。它能有效地结合其他网络安全产品的性能，对网络安全提供主动性和实时性进行全方位地保护，使用入侵检测系统（IDS）可以做到对网络边界点的数据进行实时的检测，对访问服务器的数据流进行检测，有效的发现拒绝服务攻击（Dos）等各种攻击行为，防止入侵者的破坏。

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，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，有效的降低误报及漏报率。

2　专家知识库入侵检测系统的设计

2.1　系统结构设计

一个基于规则匹配的入侵检测系统可以对多目标系统的用户活动进行实时监控的全方位入侵监控。通过数据挖掘算法分析网络数据包，与建立好的入侵系统专家的知识库匹配，从中找出正常或带有恶意的用户行为，其中入侵系统的专家知识库可以利用数据挖掘方法自动的进行更新。系统结构如图1所示。

系统包括主要包括数据采集、入侵检测（规则匹配）、决策处理、入侵响应、操作日志记录和攻击日志记录五大功能模块。根据数据流向，系统的工作流程为：

（1）数据采集模块负责从网络流量中捕获各种类型的数据帧，将这些帧交给入侵检测模块中的数据预处理模块。

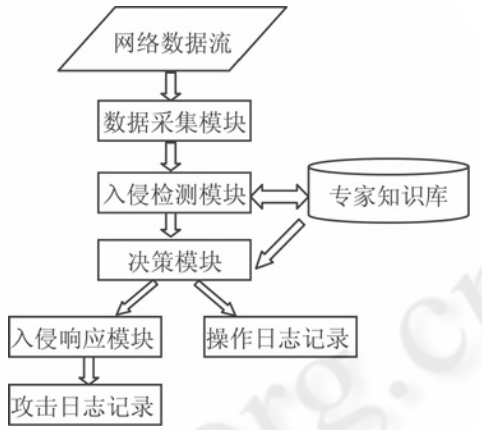

图1　专家知识库入侵检测系统

（2）数据预处理模块负责剥离帧头部，产生符合数据挖掘要求的特定格式的事件序列，

并存入事件库校验报文的完整性之后，同时将数据送给入侵检测引擎，根据规则库中现存规则对由数据挖掘算法对规则进行模式匹配，并将检测结果交给决策模块。

（3）决策模块根据专家知识库中的规则进行相似度或偏离度判断是否为入侵。

（4）入侵响应模块会及时响应，发出报警，切断连接，记录攻击日志，否则会进行一个详细的操作日志记录。

2.2　知识发现的建立过程

在基于专家知识库的入侵检测系统中，专家知识

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

库可以记录和描述用户正常和异常的行为特征并对记录已知攻击行为模式和利用已知系统漏洞进行的攻击行为模式。所以知识库的建立是非常重要的,一个好的知识库可以减少误报和漏报率。数据挖掘是专家知识库建立过程中一个特定的步骤,它通过专门的算法在网络数据中抽取模式,当网络的数据包到达入侵检测系统后,需要进行预处理,进行降噪处理,得到例如时间、源 IP、源端口、目的 IP、目的端口、连接的 ID、连接状态等重要的信息。这样的信息更有利于进行数据挖掘,符合条件加入专家知识库。系统知识库的建立过程如图 2 所示。

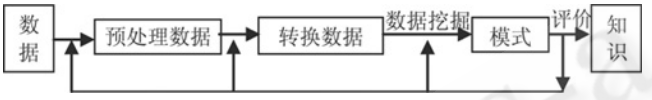

图 2　知识库发现和建立过程

对于已知的攻击行为模式和利用已知系统漏洞进行的攻击行为模式可由人工按指定的规则把这些特征加入专家知识库,这样对检测特定的入侵攻击更具有针对性和准确性,对于未知的可以通过收集足够的训练数据来训练数据挖掘模块而得到规则,利用数据挖掘把从训练数据中挖掘出的规则不断加入到专家知识库中。

3　数据挖掘分析

3.1　关联分析

在数据挖掘种关联分析是发现知识的一种重要的方法,若两个或多个数据项的取值之间重复出现且概率很高时,就存在某种关联。在知识库的建立中,关联分析采用 APRIORI 算法对数据进行挖掘。Apriori 算法它用于发现‘ 90% 的客户在购买商品 A 时也会购买商品 B ”之类的规则。下面给出关联规则中的几个基本概念。

设 $I=\{i_1,i_2,\cdots,i_n\}$ 是项的集合,设 D 为事务的集合,事务 T 是项的集合,并且 $T\subseteq I$。设 A 是 I 中的一个项集,如果 $A\subseteq T$,那么事务 T 包含 A。

定义 1 :设 A,B 是两个集合,且 $A\cap B=\Phi$ 。规则的支持度为 S、置信度为 C。支持度 S 是指 D 中事务包含 $A\cup B$(同时包含 A,B)的概率,即 $P(A\cup B)$。置信度 C 是指 D 中包含 A 的事务同时也包含 B 的概率,也就是条件概率 P(B|A)。这样关联规则可以表示为如下的蕴含形式 :$A\rightarrow B[S,C]$。

定义 2 :项的集合为项集,包含 k 个项的项集称之为 k – 项集。如果项集满足最小支持度,则它称之为频繁项集。

关联规则的目的是寻找频繁项集,Apriori 算法的重要部分就是进行剪枝,将支持度低的项集排除。而在网络攻击过程中,利用扫描工具探测系统的弱点绝大部分是攻击的一种前奏,扫描攻击中进行扫描的客户端对目标 IP 进行访问的端口的数据在所有数据包中并不多见,对于这种在收集到的数据中并不占主流的行为,如何准确的发现这种扫描攻击是入侵检测系统需要做到的一个方面。一个常见的客户端访问事务如表 1 所示。

表 1　访问事务数据

客户端	某 IP 端口列表
A	80,21
B	80
C	21
B	80
B	80,21
C	80
D	21
E	1034,1035

可以看出在对某 IP 不同端口进行访问的事务中,80 端口访问的 s = 62. 5% ,21 端口访问的 s = 50% ,而 1034 和 1035 访问的支持度仅为 12. 5% ,置信度为 100% 。但这很可能就是一种进行类似扫描行为的攻击。

在此,可以对 Apriori 算法进行适当的变化,主要是对产生新后候选集函数的修改,改进后的算法为 :

```
Procedure apriori_gen(L_{k-1},min_sup)
  C_k = Φ ;
  C_k' = Φ ;
    For each itemset li ∈ L_{k-1}
      For each itemset lj ∈ L_{k-1}
      If (l_i[1] = l_j[1]) ∧ (l_i[2] = l_j[2]) ∧ ⋯ ∧
(l_i[k-1] = l_j[k-1]) then
        {  c = l_i join l_j ;
          if has_omfreqient_subset(c, L_{k-1}) then
```


©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

add c to C_k ;//产生非频繁集

else add c to C_k ' ;//频繁集

}

return C_k ;

C_k '可以进行通过其他的数据挖掘方法重新进行挖掘,C_k 为本次挖掘所要寻找的项集。对于这类攻击行为,训练样本和支持度的选择很重要,支持度太高,会产生误判,太低很可能收集不到样本。

3.2 序列模式在 Apriori 算法的应用

在网络攻击行为中,试图对某 IP 进行扫描攻击类似的异常行为可以通过上述方法获得,但对于其他很多攻击行为,例如 Dos 攻击却无能为力。Dos 攻击是一类主要的网络攻击方法,Dos 攻击的特征在于在一段时间内大量的发送数据包,常用的方法有 SYS Flood。SYS Flood 在一段时间内伪造大量的虚假地址对某 IP 发起攻击,导致服务器大量的资源长期处于半连接的等待状态,最终资源耗尽。

对 Dos 攻击进行关联分析时,如果按每个 IP 对某服务器 IP 发起的访问为一个事务,挖掘出来发现支持度与置信度比较低,所以会造成漏判。所以需要进行改进,为了提高检测速度,可以用 C k '对数据进一步进行挖掘。

Apriori 算法的关联挖掘的是形如 A→B[S,C] 的规则, Apriori 新算法在处理过程中,应该把记录的每个属性值对看成是一个事务,数据一般需要进行丢弃、离散化和格式转换等数据预处理,这几个连续变量进行离散化,以便于关联规则的发现。在这里取 5 个属性用于数据挖掘(time,sip,protocol,desip,state),记录的属性主要包括有时间,源地址,协议,目标地址、状态。将与所有事务按时间戳(此处的时间戳为一设定的单位时间),可以得到对某 IP 访问的一个序列,如果状态为等待连接或拒绝时,可以从中发现 Dos 的频繁项集。

通过序列模式的类 Apriori 的算法主要找出在单位时间戳内同时连接状态为拒绝或等待的频繁序列。序列模式发现的类算法为:

L1 = {i | i ∈ I ∧ P(i)/N > = minsup} ;

For (k = 2 ;$L_{k-1} \neq \Phi$;k + +)

{C_k = apriori - gen(L_{k-1}) ;//C_k 是 k 个元素的候选集

FOR all transactions t ∈ D DO

{C_t = subset(C_k ', t) ;//C_t 是所有 t 包含的候选集元素

FOR all candidates c ∈ C_t

c. count + + ;

}

L_k = {c ∈ C_k ' | c. count > = minsup_count}

}

上述算法几乎和 Apriori 一样。该算法将迭代产生新的候选 l - 序列,剪掉那些(l - 1) - 序列的非频繁序列的候选,然后对留下的候选计数,识别序列模式。

4 实验结果

在实验过程中,在数据库中提取了 10000 条记录作为样本数据进行训练,其中包括有对 2 个常用服务器 IP 地址的访问数据,在对扫描攻击时产生的 s = 36.86% ,c = 82.32% 。从中可以看出扫描攻击的可能性很高,而且随着训练数据加到 15000 条,s 的值降低到了 23.22% 。

同样在进行 Dos 的检测测试中,通过样本数据,在单位时间内的 s 达到了 96.28% ,c = 93.32% 。

5 结束语

随着数据挖掘技术特别是挖掘算法研究的深入,给优化系统的性能提供了途径,同时充分溶合了异常检测技术和误用检测技术,实现了两种技术的优势互补,提高了系统的可扩展性、适应性和准确性。

参考文献

1 K. Ali, S. Manganaris, and R. Srikant. Partial Classification using Association Rules. In Proc. of the 3rd Intl. Conf. on Knowledge Discovery and Data Mining, pages 115 - 118, NewPort Beach, CA, August 1997.

2 C. C Aggarwal and P. S. Yu. Mining Large Itemsets for Association Rules. Data Engineering Bulletin, 1998, 21 (1): 23 - 31.

3 Han J, et a1. Data Mining: Concepts and Techniques [M]. 北京:机械工业出版社,2001:149 - 222.

4 Margaret, H - Dunham. Data Mining: Introductory and Ad. vancedTopics[M]. 北京:清华大学出版社,2003: 164 - 192.

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn