基于关联规则挖掘的分布式入侵检测系统研究①

Distributed Intrusion Detection Systems Based on Mining Association Rules

费洪晓　黄勤径　（中南大学信息科学与工程学院 长沙 410075）
谢文彪　（中南大学信息科学与工程学院 长沙 410075）
（长沙理工大学电气与信息工程学院 长沙 410076）
戴　弋　（中南大学信息科学与工程学院 长沙 410075）

摘要：针对现有应用数据挖掘技术的入侵检测系统存在实时性差、难以提取有效的特征属性、漏报和误报率较高等问题，论文设计了一种基于关联规则挖掘的分布式网络入侵检测模型，阐述了如何从原始审计数据中提取和构造属性集，并将模糊逻辑和增量更新技术结合以提高系统的检测效率、准确性和自适应能力。试验证明了该系统实现的可行性。

关键词：关联规则　分布式入侵检测　模糊数据挖掘　增量更新

1　引言

随着大规模网络入侵事件的发生，入侵检测技术越来越受到人们的关注。IDS（Intrusion Detection Systems，入侵检测系统）是指对于面向计算机资源和网络资源的恶意行为识别和响应的网络安全系统，是包括技术、人、工具三方面的一个整体。然而随着攻击技术、方法与攻击工具的变化和发展，传统的依靠经验方式建立的基于专家系统的 IDS 在扩展性、适应性以及分布式攻击检测和数据协同分析等方面已经暴露出明显的不足。同时分布式入侵检测因其具有良好的分布式攻击检测能力、开放性架构和较强的容错和抗毁能力，从而成为当前网络安全研究领域的热点[1]。

数据挖掘本身是一项通用的知识发现技术，其目的是要从海量数据中提取出我们所感兴趣的数据信息（知识）。将数据挖掘技术应用到入侵检测领域，构建基于数据挖掘的入侵检测系统，其主要特点是：它能够自动地从大量的网络数据中构建简洁、精确的正常的或入侵行为模式，解决了传统入侵检测系统手工分析以及对攻击模式进行硬编码的沉重负担。由于该方案具有通用性（它可以处理各种结构化的数据）和自动处理的功能，因而可在许多不同的网络环境中构建相应的入侵检测系统[2]。

本文设计了一个基于关联分析算法的分布式入侵检测系统原型，从系统结构和挖掘算法方面提出了自己观点，目的是在异构网络环境中、在攻击手段改变的情况下，对来自于分散的监控点的数据进行综合的基于关联规则的数据挖掘分析，实时发现网络攻击行为。

2　系统原理与模块结构

系统原型采用分布式结构，其结构模型如图 1，包含有：对网络数据的数据采集、数据准备和预处理、特征变量选取、基于误用和异常模式的检测、模式的增量更新和分发等一系列的过程。系统既可对同一局域网实时监控，也可监控不同网络，具有良好的伸缩性和扩展性。对于大型网络的分布式入侵检测系统，数据采集和预处理、误用和异常行为检测由分布在各个监控点的主机完成，自适应模式挖掘和数据仓库在后台服务器执行，组件之间利用 IBM Aglets 的消息 Message 对象实现相互之间的通信。

（1）数据采集和预处理。采用改进的基于 Linux 内核的 Bro 作为网络数据包过滤及重组引擎，通过调用 libpcap 接口从链路层获取数据帧，进行逐层的协议解

① 基金项目：国家自然科学基金面上项目（60673165）；湖南省自然科学基金（05JJ30119）；湖南省科技计划项目（2006JT1040）。

析，恢复成基于传输层或更高层的连接记录。这些连接记录是网络活动（如 HTTP、FTP 记录）或者用户网络行为（如 TELNET 记录）的基本描述。我们可以从中提取有意义的特征属性，例如一条 TCP 连接可以用以下属性表示：

<时戳、耗时（秒）、源 IP、目的 IP、源端口、目的端口、字节数、协议类型、连接结束状态>

审计数据收集后，在数据挖掘之前还需进行量化属性的离散化工作，如源地址，目的地址等量化属性可以借助原有系统的拓扑发现，按照子网进行离散化，从而达到有效降低挖掘工作量的目的。

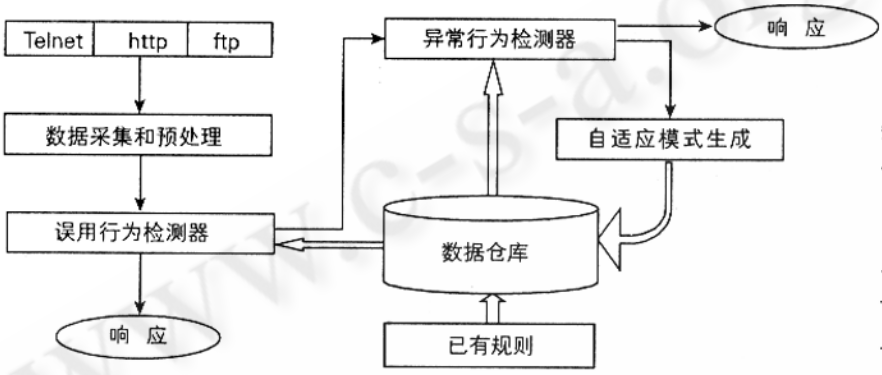

图 1　入侵检测系统结构模型

（2）误用行为的挖掘。数据采集和预处理模块将网络中传输的数据包还原成基于传输层的连接记录，从中提取出可以对连接记录进行分类的特征属性。对于在传输层无法判断的连接记录，则进行高层的协议解析，分解成相应的 Ftp、Telnet、Http 会话，针对每一种高层协议，提取出可以用于判断的特征属性，并送交误用行为检测器模块采用模式相似度匹配来检测常规入侵，一旦发现误用行为即激活响应模块。误用行为模式库保存在数据仓库中，方便对于规则库的实时更新。

（3）异常行为的挖掘。异常检测主要针对于 Telnet 会话过程中用户所执行的 shell 命令集的挖掘。将误用行为检测器未能检测出异常的网络审计数据提交异常行为检测器，利用相似度函数进行分析，如发现异常即激活响应模块。同时将检测结果送入正常或异常数据集，同时可作更新误用行为模式库或用户正常行为库之用。

（4）自适应模式生成器和数据仓库。模式生成器和数据仓库是整个系统的核心部分，运行于后台服务器内。入侵检测系统的检测性能很大程度上决定于误用行为模式库的完备性以及用户正常行为模式的准确性和实时性。模式生成器的任务就是生成原始的正常和异常行为模式，并尽可能早的发现和分发新的入侵检测模型。首先将系统运行期间误用和异常行为监测器未能检测出攻击的不同监控点主机审计数据的量化属性进行进一步分段或抽象成更高的概念级，再对其进行模式挖掘，并对正常网络活动或用户行为模式的改变以及新的异常行为模式进行发布。

数据仓库是数据和模式的中心存储器。将所有数据集中存储的优点是不同组件可以异步处理数据库中的同一条记录，例如离线的数据训练或人工修改规则等。同类组件（如各监控点主机）也可以同时调用数据仓库中的数据。关系数据库特征支持“存储过程调用”，可以在后台服务器上高效率的自动运行各种复杂计算。任意数量的监控点数据都可以由一条简单的 SQL 语句在数据仓库中找到，同时利用数据仓库也可以很好的综合分析不同监控点或不同网络所收集的数据，支持数据共享和协同分析，使复杂的分布式攻击的检测成为可能。

3　提取和构造属性集

入侵检测的两个基本前提是系统活动能被观察，比如说经由审计，并且有明显的“证据”来区分正常和入侵行为。这种从原始审计数据中提取的“证据”即属性（feature），我们运用这些属性来建立和评估入侵检测模型。属性提取实际上就是一个决定从原始审计数据提取对于分析入侵最重要的何种“证据”的过程，因此，属性集提取和构造在建立入侵检测系统中是非常关键的一步[3]。

系统采用由 Wenke Lee 提出的从审计数据中选取和构造属性集的数据挖掘算法。首先，原始审计数据（二进制）被处理和归纳成离散记录，包括大量的基本属性，如一条 TCP 连接可以用以下属性表示：时间戳，耗时，源 IP，目的 IP，端口号和错误标志等。专门的数据挖掘程序将被运用到这些记录中来计算频繁模式，描述记录中属性和频繁发生事件之间的相关性。一个

基本的模式是 A:B - >C:D[置信度,支持度],即 A 和 B 在某种置信度下可以推导出 C 和 D,并以某种几率(模式支持度)发生。然后通过对正常行为模式与某种入侵相关的"独特"行为模式的鉴别和分析,来为连接记录构建附加属性。事实表明,构造属性能够把正常记录和入侵记录清晰的区分开来。运用这个方法,构造属性是以实际经验数据为基础,因此比专家知识更加客观。具体的算法见文献[4]。

下面以一次 SYN flood 攻击为例来说明此过程。当发起这种攻击时,攻击者使用很多伪造的源地址来建立与被攻击主机某个端口的连接(如 http),而这些连接是不可能被完全建立的(即只有第一个 SYN 包被发送了,连接仍然是处于"S0"状态)。通过分析比较包含了 SYN flood 攻击的数据集和来自同一网段的正常数据集这两者的模式,计算出二者的差异。我们可以得到以下规则:(flag = S0, service = http, dst_host = victim),(flag = S0, service = http, dst_host = victim) → (flag = S0, service = http, dst_host = victim) [0.93, 0.03, 2]。这表明在与被攻击主机主机进行两次 S0 标志的 http 连接后的 93% 时间里,在前面两个连接发生后的 2 秒内,将进行第三个类似的连接,这种模式发生的几率是总数据的 3%。从而,属性构造算法可以解析模式属性:"统计在 2 秒内对同一目标主机进行连接的数量来计算连接中具有相同服务的几率和具有相同 S0 标志的几率"。对于这两个"百分率"属性,正常连接记录的值接近 0,但是 SYN flood 的连接记录的值却高于 80%。一旦这些有判别能力的属性被构建了,就很容易通过手动(即人工编码)或者自动(即机器学习)技术产生入侵规则。例如:利用提取到的属性计算出 SYN flood 入侵规则:如果在 2 秒中内,对同一主机的连接记录超过 4 次,使用同一服务的几率超过 75%,并且具有"S0"标志的几率大于 75%,那么这就是一次 SYN flood 攻击。

4 算法设计

关联规则挖掘的目标是发现每条审计记录内部不同属性之间的相互依赖的模式。关联规则定义如下:定义 $I=\{i_1,i_2,\cdots,i_m\}$ 为项目全集,D 为事务数据库,其中每个事务 T 是一项目子集($T \subset I$),并具有唯一的标志符 ID。关联规则是形如 $X \rightarrow Y_{[c,s]}$ 的逻辑蕴含式,其中 $X \subset T, Y \subset T, X \cap Y = \Phi$,s 是 $X \cup Y$ 的支持度(表中同时包含 X 和 Y 的记录所占的百分数),c 是该规则的置信度,定义为 $s_{X \cup Y}/s_X$。

本文对审计数据关联规则的产生结合了模糊逻辑和增量更新算法,首先利用模糊关联规则产生频繁项集,然后在此基础上增量更新。

4.1 模糊数据挖掘算法

审计记录中包含有量化特征,量化数据在挖掘过程中由支持度和置信度的阈值分隔在两个区间中。这种分区带来的所谓"尖锐边界"问题会使 IDS 产生更高的漏报率和误报率[5]。本文通过将模糊逻辑和关联规则算法结合产生频繁项集,较好的解决了这个问题。

模糊关联规则和关联规则类似,不过,项的全集 $I=\{i_1,i_2,\cdots,i_m\}$ 中的每个元素不再是经典的概念,而是模糊概念了,事件 $t=<t_1,t_2,\cdots,t_m>$ 中的元素 t_k 也变为对 i_k 的隶属度了。

由于基本模糊规则挖掘算法没有考虑到入侵检测领域知识,经过挖掘会产生大量的"无趣规则",所以在算法中引入对行为模式描述较为有用的属性约束规范,有助于提高挖掘算法的效率。为了避免约束规范的引入而遗漏某些重要的频繁项集,我们对属性约束定义如下:审计记录中的属性集,按照在入侵检测中的意义不同分为三个集合,核心属性集 $A=\{a_1,a_2,\cdots,a_n\}$,条件属性集 $P=\{p_1,p_2,\cdots,p_n\}$,结果属性集 $R=\{r_1,r_2,\cdots,r_n\}$。约束条件表示为:$p_1 \wedge p_2 \wedge \cdots \wedge p_l \rightarrow r_1 \wedge r_2 \wedge \cdots \wedge r_k$,其中 $p_n(1 \leqslant n \leqslant l)$、$r_m(1 \leqslant m \leqslant k)$ 且满足 $p_n \in P$、$r_m \in R$,且至少存在一个属性 $p_i \in A(1 \leqslant i \leqslant l)$。

4.2 增量更新算法

4.1 节介绍了一种从数据库挖掘频繁项目集的基于属性约束的模糊挖掘算法,但是对于一个实时入侵检测系统,每天都要向数据库中增加大量新的审计数据,须对其进行增量更新挖掘。

对于给定事务数据库 D 和一个新的事务数据集 d,假定最小支持度不变,事务数据库 D 中增加了新的事务数据集 d 后,最新事务数据库 $D \cup d$ 中关联规则的高效更新问题,一种可能的方法就是将关联规则的挖掘算法如 Apriori 对最新事务数据库 $D \cup d$ 重新运行一遍,这种方法虽然简单明了,却有着明显的不足,因为最初用来发现旧的频繁集的计算都将被浪费,所有的频繁项目集都必须从头开始计算。为此,本文提出了

一种新的增量关联规则更新算法，该算法只需扫描原库一次，大大减少了运算量。

根据频繁项目集的定义可以知道：对于任意一个X项集，若其在D和d中都是频繁项集，则在D∪d中同样是频繁项集，其支持数为二者之和；若其只在D中是频繁项集，则其支持数应该加上d中的支持数以决定它在D∪d中是否为频繁项集；若其只在d中是频繁项集，则其支持数应该加上D中的支持数以决定它在D∪d中是否为频繁项集；若其在D和d中都是非频繁项集，则其在D∪d中就是非频繁项集。

文献[6]提出了一种通用的关联规则增量式更新算法，本文对其进行改进，结合模糊挖掘，将新事务集d转换成模糊数集，利用上面提到的频繁项集的性质，对频繁项集进行更新。其具体算法如下：

```
Begin
(1) 将 d 转换成 d;
(2) 利用模糊数据挖掘方法生成 d 的频繁项目集 L1;
(3) For(x∈I) //I 为全集,包含所有的事务集
{
If(x∈L&&c∈L1) //L 为原事物库 D 的频繁项集
  { D1 =L∩L1;}
If(x∈L&&x 不属于 L1)
  {
   D2 =L - L∩L1; //项 x 在 D 中频繁,在 d 中不频繁
   扫描 d,读出 x 的 count; //count 为模糊关联挖掘时对每个项进行的支持度计算结果
   Support(x,D∪d) = Support(x,D) + Support(x,d);//计算 x 在 D∪d 中的支持度
   If(Support(x,D∪d) <min_support)
   D2 = D2 - {x}; //删除支持度小于最小支持度的项
  }
If(x∈L1 &&x 不属于 L)
  {
   D3 = L1 - L∩L1; //项 x 在 d 中频繁,在 D 中不频繁
   扫描 D,读出 x 的 count;
   Support(x,D∪d) = Support(x,d) + Support(x,D);
   If(Support(x,D∪d) <min_support)
   D3 = D3 - {x};
  }
 L2 = D1∪D2∪D3; //生成 D+d 的频繁项集
 }
End
```

5 结束语

系统性能是衡量一个入侵检测系统是否实用的重要标准。本系统实现分布式设计框架，数据采集、预处理以及误用和异常模式相似度的匹配由各监控点主机完成，数据仓库和模式的生成放在系统服务器，有效的减少了各监控点主机的资源占用，提高了检测的实时性和检测效率，并且有利于系统的集中管理和协同工作。检测过程中，数据经过了三层过滤：数据预处理模块对审计数据的误用检测和异常检测屏蔽了大部分的网络入侵；为自适应模式生成器提供了尽可能精简的数据集，提高了数据挖掘算法的速度，减少了监控点主机和服务器之间的通信量；同时模糊逻辑和增量更新技术的结合运用有效的提高了系统的检测效率、准确性和自适应能力。

参考文献

1 罗守山，入侵检测[M]，北京：北京邮电大学出版社，2005.15～47.

2 卿斯汉、蒋建春、马恒太，入侵检测技术研究综述[J]，通信学报，2004，25(7) 19～29.

3 W. Lee, S. J. Stolfo, P. K. Chan, E. Eskin. Real Time Data Mining-based Intrusion Detection[J]. 2001,1(1): 89～100.

4 W. Lee, S. J. Stolfo, K. W. Mok. Algorithms for mining audit data. In T. Y. Lin, editor, Granular Computing and Data Mining. Springer-Verlag, 2000. to appear.

5 向继东，基于数据挖掘的自适应入侵检测建模研究[D]：[博士学位论文]. 武汉：武汉大学，2004.

6 薛锦、陈原斌，一种实用的关联规则增量式更新算法[J]，计算机工程与应用，2003，39(13)，212～214.