基于数据流的网络入侵检测研究

朱桂宏,王 刚
(铜陵学院 计算机系,安徽 铜陵 244000)

摘 要:目前,入侵检测系统面临数据量大、内存等系统资源不足的问题。将两阶段聚类算法应用于入侵检测,设计了基于数据流的入侵检测系统模型。实验结果表明,该系统可以取得较高的检测率和较低的误报率,具有自适应性和可扩展性,并有效降低了对内存资源的需求。

关键词:入侵检测;数据流;聚类分析
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2009)03-0175-03

Research on Network Intrusion Detection Based on Data Stream

ZHU Gui-hong, WANG Gang
(Department of Computer Science, Tongling College, Tongling 244000, China)

Abstract: At present, intrusion detection system faces some problems: a large quantity of data, memory lacking, and so on. Applying a two-stage clustering algorithm to the intrusion detection, a IDS model based on data stream is designed. The empirical results manifest that the system can achieve a higher detection rate and a low rate of false positives, with adaptability and scalability, and effectively reduce the demand for memory.

Key words: intrusion detection; data stream; clustering analysis

0 引 言

随着 Internet 的快速发展与日益普及,越来越多的信息通过网络来传输和存储,网络安全显然越来越重要。目前常见的网络安全技术主要有加密、数字签名、身份认证、访问控制、防火墙技术和入侵检测技术等。

入侵检测(Intrution Detection)是通过监视系统或网络流量、系统审计记录等来发现与识别对系统和网络的入侵企图和入侵行为[1]。入侵检测是防火墙的合理补充,帮助系统对付网络攻击,它扩展了系统管理员的安全管理能力,提高了信息安全基础结构的完整性。网络入侵检测是以截获的网络数据包为信息源进行的入侵检测。

入侵检测方法可以分成滥用检测(Misuse Detection)和异常检测(Anomaly Detection)。滥用检测方法是将已知攻击的攻击特征保存在特征库中,通过模式匹配的方式检测攻击。其优点是可以有效地检测出已知的攻击方式,误报率低;缺点是不能检测出未知的攻击。异常检测是建立正常行为模式,通过正常行为与系统或网络行为之间的差异来检测攻击。其优点是可以检测到新的未知攻击,缺点是误报率较高。

当前的基于数据挖掘的网络入侵检测方法通常要对所有的网络数据包进行挖掘,时间和空间复杂度较高,且容易受到内存资源的限制。在线检测过程中,要处理大量、持续到达的网络数据包,用现有的检测方法难以及时、准确地进行检测。因此,提出了基于数据流的网络入侵检测模型。

1 数据流

数据流就是大量连续到达的、潜在无限的数据的有限集合。令 t 表示任一时间戳,a_i 表示在该时间戳到达的数据,数据流可以表示为 $\{\cdots, a_{i-1}, a_i, a_{i+1}, \cdots\}$。与传统的数据相比,数据流有许多自己的特点[2-4]:

①数据高速到达,实时性要求高。数据流的快速流动性要求挖掘算法对数据流的分析处理速度不能低于数据流的流动速度。

②数据流是一种大量、快速到达的数据,数据容量将很快超过内存或硬盘的存储量,因此不可能对数据流中的每一个数据都进行存储。

③由于数据量无限增长,对数据流的扫描次数仅

收稿日期:2008-07-08
基金项目:安徽省自然科学研究项目(KJ2008B41ZC)
作者简介:朱桂宏(1977-),男,安徽铜陵人,硕士,讲师,研究方向为网络安全、数据挖掘。

限于一次,即数据一经处理,除非刻意保存,否则不能被再次取出处理,或者再次取出代价昂贵;而在传统数据挖掘中,数据集通常保存在数据文件中,可对数据集进行多次扫描。

④数据流数据量的无限性使得数据流挖掘无法保存原始数据,仅能在内存中保留原始数据的概要信息,并基于这些概要信息生成最终结果。因此,数据流挖掘结果实际上是在一定误差范围内的近似结果。

由于网络通信量与数据流有着天然的联系,符合以上的数据流特点,所以采用数据流模型来描述实际的网络通信量,解决现有入侵检测模型存在的不足是非常合适的。

2 基于数据流聚类分析的入侵检测模型

2.1 系统模型

系统模型如图 1 所示。

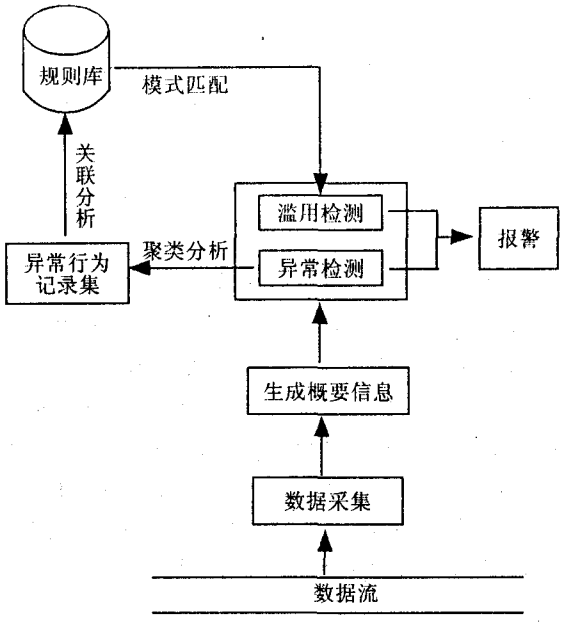

图 1 基于数据流的网络入侵检测系统模型

2.2 数据采集模块

数据采集模块负责抓取来自 Internel 的网络数据包,通过协议分析,生成系统需要的网络连接记录。网卡一般有两种接收方式:混杂模式和非混杂模式。在混杂模式下,不管数据帧中的目的地址是否与自己的地址匹配,都接收下来;在非混杂模式下,只接收与目的地址相匹配的数据帧及广播数据包(包括组播数据包)。为进行网络监听,网卡必须被设置为混杂模式。

系统使用 Winpcap 获取数据包,并对其进行处理,提取包头信息,生成网络连接记录。

2.3 生成概要信息模块

在概要信息生成模块中,通过对不断到达的网络数据流进行单次扫描聚类,生成了描述原始网络数据流的概要信息[5]。概要信息用聚类得到的子聚类及其特征值 CF 来表示。CF 是一个 7 元组($\overrightarrow{CF^2}$,$\overrightarrow{CF^1}$,$\overrightarrow{CF^0}$,t^2,t,d_{begin},span),其中$\overrightarrow{CF^2}$、$\overrightarrow{CF^1}$、$\overrightarrow{CF^0}$ 分别对应于子聚类中的点集数目、算术和以及各维向量的平方和,t、t^2 表示数据点到达的时间的和及平方和,d_{begin} 表示该子聚类的起始点,span 表示该子聚类的体积,$\overrightarrow{CF^1}/\overrightarrow{CF^0}$ 用来表示该子聚类的聚类中心。

事实上,用户对于最近的数据更感兴趣。因此,只需要对少量的近期数据进行细节分析,而对大量的历史数据,仅给出一个概要。在存储概要信息时,采用了基于时间窗口模型的金字塔时间框架(pyramid time frame)[6]的结构。这样,只需要一个较小的数据窗口,就可以存储概要信息,大大减少了系统对内存的需求。

系统模型中的概要信息生成算法如图 2 所示。

```
算法:概要信息生成算法
输入:代表网络连接记录的数据流,滑动窗口的宽度
w,预先定义的数据块内密度 Dcin
输出:描述原始网络数据流的概要信息
for (i=0;i<w;i++)
  读入一个数据点;
  if (该点已属于某个内存中已存在的数据块)
  修改该数据块对应的聚类特征值;
  else
  为该点创建一个聚类块;
  endif
  重新计算各聚类块的密度函数 Dcin;
  扫描内存中所有的聚类块
  {if (某数据块的密度<=Dcin)
    if (块内密度<=可抛弃的密度)
      抛弃该数据块;
    else
      压缩聚类块;
    endif
  else
    聚合相似、相邻的数据块;
  endif
}
按 paramid time frame 结构存储数据;
```

图 2 概要信息生成算法

2.4 入侵检测模块

在入侵检测模块,系统首先对概要信息利用 k-means 算法进行聚类分析,生成正常和异常用户行为记录集。再对异常用户行为记录集采用 Aprior 算法进

行关联分析,挖掘出入侵行为模式,产生入侵检测规则,并自动加入规则库中。最后,将输入的网络连接记录与规则库中的规则进行模式匹配,若匹配,就进行报警。

3 实验及分析

3.1 实验环境

(1)硬件环境:Intel Pentium 2.4GHz CPU,512MB内存。

(2)软件环境:windows 2000 操作系统,Microsoft Visual C++ 6.0。

(3)实验数据集:文中选用 KDD CUP'99 数据集[7]作为实验数据集。此数据集中包含了500万条网络连接记录,每条网络连接记录有7个分类属性和34个数值属性。

数据集包含4种主要的攻击类型:

DoS——拒绝服务攻击;

Probe——扫描与探测行为;

R2L——对远程主机的未授权的访问;

U2R——对超级用户权限的未授权的访问。

3.2 聚类质量

聚类质量采用聚类分析中最常用的距离平方和(sum of squared distance,简称SSQ)作为度量指标,具体定义[6]如下:设当前时标为 T_c,窗口大小为 N,对各个处于 T_c-N 到 T_c 间的数据点 p_i,Cp_i 是数据点 p_i 所属的子聚类的中心,然后计算 p_i 和 Cp_i 之间的距离 $\mathrm{dist}(p_i,Cp_i)$;最后计算在时刻 T_c 窗口大小为 N 的 SSQ,$\mathrm{SSQ}=\sum_{N}[\mathrm{dist}(p_i,Cp_i)]^2$。

将文中用到的两阶段聚类算法 Sdstream 与 Stream 算法进行比较,分别计算它们的 SSQ 值。如图3所示。

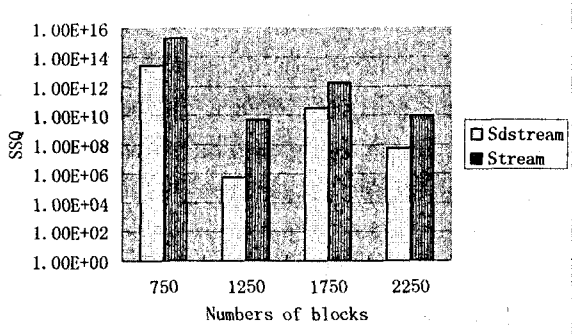

图3 Sdsteam 与 Stream 的 SSQ 值比较

3.3 入侵检测结果

文中利用 KDD CUP'99 网络入侵检测数据集对系统进行了测试,结果如表1所示。实验表明,系统对DoS攻击具有很高的检测率,而对其它类型的攻击检测效果略差。

表1 入侵检测结果

攻击类型	检测率(%)	误报率(%)
DoS	98.31	1.93
Probe	83.63	0.31
R2L	63.71	0.23
U2R	65.84	0.87

4 结束语

入侵检测是网络安全技术领域的主要研究方向之一。将数据流挖掘技术应用到入侵检测系统中,可以自动地从大量的网络数据中发现新的模式,减少手工编写入侵行为模式和正常行为模式的工作量。

文中将基于数据流的两阶段聚类算法应用于入侵检测,系统能够有效地建立网络正常行为模型,并且显著提高了入侵检测的速度。同时,通过关联分析能够从用户异常行为记录集中挖掘出新的入侵行为模式,并自动产生入侵检测规则,使得系统具备检测新类型攻击的能力。

对数据流挖掘技术结合到网络入侵检测系统的研究只是一个初步尝试,今后将在现有的基础上继续作深入的研究。

参考文献:

[1] 唐正军,李建华.入侵检测技术[M].北京:清华大学出版社,2004.

[2] O'Callaghan L. Approximation algorithms for clustering streams and large datasets[D]. California: The Department of Computer Science, Stanford University, 2003.

[3] Muthukrishnan S. Data Streams: Algorithms and Applications [C/OL]//Proceedings of the 14th Annual ACM - SIAM Symposium on Discrete Algorithms. 2003. http://athos.rutgers.edu/~muthu/stream-1-1.ps.

[4] 金澈清,钱卫宁,周傲英.流数据分析与管理综述[J].软件学报,2004,15(8):1172-1181.

[5] 朱蔚恒,印鉴,谢益煌.基于数据流的任意形状聚类算法[J].软件学报,2006,17(3):379-387.

[6] Aggarwal C, Han J, Wang J, et al. A Framework for Clustering Evolving Data Streams [C]//Proceedings of the 29th VLDB Conference. Berlin, Germany: [s.n.], 2003.

[7] KDD99 CUP dataset [EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.