基于熵的流量分析和异常检测技术研究与实现

崔锡鑫,苏 伟,刘 颖
(北京交通大学 电子信息工程学院,北京 100044)

摘 要:随着互联网的飞速发展,网络安全问题受到越来越多的关注。作为一种重要的网络监管控制手段,流量异常检测技术也越来越受到人们的重视。目前流量异常检测方法有很多,基于熵的流量异常检测是近几年研究较多的一种方法。文中在基于熵的流量异常检测的基础上,先对两种算法进行编程实现,即基于信息熵的流量异常检测算法和基于联合熵的流量异常检测算法,而后对这两种算法进行实验测试与分析比较,结果表明基于联合熵的流量异常检测可以更为有效地检测出异常。同时根据分析结果,提出一种有效的检测流量异常的分析思路。

关键词:异常检测;熵;联合熵
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2013)05-0120-04
doi:10.3969/j.issn.1673-629X.2013.05.031

Research and Implementation of Traffic Analysis and Anomaly Detection Technology Based on Entropy

CUI Xi-xin, SU Wei, LIU Ying
(College of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: With the rapid development of Internet, the network security receives more and more attention. As one of the important means of network monitoring and control, the traffic anomaly detection technology has been realizing more important by people. Now there are many ways to detect the anomaly of the traffic, and the anomaly detection technology based on entropy is an important means in recent years. First the anomaly detection technology based on information entropy and the anomaly detection technology based on joint-entropy are programming realized in this paper, then make an experimental test and an analysis according to the two ways. It shows that the anomaly detection technology based on joint-entropy is more effective, and then a valid analysis of ideas to detect the anomaly of traffic is proposed.

Key words: anomaly detection; entropy; joint-entropy

0 引 言

如今流量异常检测技术[1~3]作为一种网络监管控制中的重要手段越来越受到人们的重视。流量异常检测技术源于网络安全中的入侵检测领域,Denning 于 1986 年提出了用于安全事件检测的“入侵检测模型”[4,5],该模型主要由三部分构成:信息源、分析引擎和响应组件。与入侵检测一样,流量异常检测首先在对信息源建模分析的基础上,勾画出检测对象的行为模式轮廓,通过新数据样本和行为模式轮廓的对比来发现当前行为特征的偏离。

流量异常的检测方法有很多,基于熵的流量异常检测是近几年研究较多的一种方法。熵最早是 1856 年德国科学家克劳修斯将其引入到热力学第二定律中,而后信息论的创始人香农用熵作为随机事件不确定性的度量,从而让熵从物理学走到信息学[6]。很多研究表明网络流量具有自相似、长相关和重尾分布等分布特征[7],这些发现对于网络流量工程、网络建模和异常检测具有指导意义,熵值恰好可以用来描述这些特征,从而分析信息流量特征的变化来检测流量异常。文中对两种基于熵的流量异常检测算法:基于信息熵的流量异常检测和基于联合熵的流量异常检测进行编程实现,同时进行了实验测试和数据分析,最后提出了一种有效的检测流量异常的分析思路。

1 流量异常检测技术概述

在对流量进行分析比较时,首先需要对流量信息

收稿日期:2012-07-23;**修回日期**:2012-10-26
基金项目:“新一代宽带无线移动通信网”重大专项(2011ZX03002-005-03);国家自然科学基金资助项目(61202428, 60870015, 60903150);北京市自然科学基金项目(4122060)
作者简介:崔锡鑫(1990-),男,硕士研究生,研究方向为计算机网络安全和下一代互联网;苏 伟,博士,副教授,研究方向为新一代信息网络理论与关键技术。

进行采集,因此一个完整的流量异常检测系统分为两个功能模块:流量采集模块和流量分析模块。

1.1　流量采集模块

文中流量采集模块采用基于 NetFlow[8~11] 的流量采集技术。流量采集的对象是指产生 NetFlow 流记录的网络设备。数据采集部分由采集器、收集器以及数据库组成,其中采集器负责收集、分析流入路由器或交换机的 IP 数据流,然后按需选择模版将流信息组成 NetFlow 报文发送给收集器;收集器负责接收并存储采集器发过来的 NetFlow 报文,从中提取出流记录信息存储到数据库中。

采集器可以部署在任何产生流量的地方,当系统规模较大时,需要配置多个采集器。采集系统模型如图 1 所示。

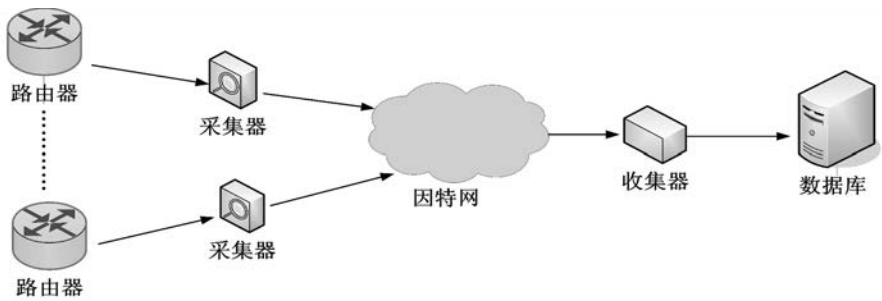

图 1　流量采集系统模型

1.2　流量分析模块

文中流量分析模块采用基于熵的流量异常检测算法。选取了两种基于熵的流量异常检测算法进行流量分析,即基于信息熵的算法和基于联合熵的算法。

2　流量异常检测算法实现

2.1　采集模块功能实现

将采集到的流量信息存储到数据库表 receive_pkt_v4 中,数据库的记录格式如表 1 所示,其中各列记录了进行流量分析所需要的信息。

表 1　数据库中流记录格式

列名	类型	约束条件	说明
no	INT	NOT NULL	数据序列号
receive_time	DATETIME	NOT NULL	处理设备接收时间
src_ip	VARCHAR(50)	NOT NULL	源地址
dst_ip	VARCHAR(50)	NOT NULL	目的地址
protocol	SMALLINT	NOT NULL	协议号
src_port	SMALLINT	NOT NULL	源端口
dst_port	SMALLINT	NOT NULL	目的端口
start_time	DATETIME	NOT NULL	数据流开始时间
end_time	DATETIME	NOT NULL	数据流结束时间
octet_count	INT	NOT NULL	数据流总字节数
pkt_count	INT	NOT NULL	数据流总包数
domain	SMALLINT	NOT NULL	采集设备 ID

2.2　分析模块功能实现

流量分析模块是对采集到的流量信息进行统计分析而后计算出熵值,根据熵值的变化检测异常。文中选取了两种策略进行流量异常检测,即基于信息熵的流量异常检测和基于联合熵的流量异常检测。

1)基于信息熵的流量异常检测。

经过流量采集得到表 1 中的信息,包括数据序列号、处理设备接收时间等。文中选取了其中的源 ip、目的 ip、源端口、目的端口作为数据处理分析并计算信息熵的对象。信息熵[12]的计算如公式(1)所示:

$$H[X] = H(p_1, p_2, \cdots, p_n) = -k\sum_{i=1}^{n} p_i \log p_i \qquad (1)$$

以源 ip 的信息熵计算为例,$p_i = n_i/s$,n_i 对应的是第 i 个源 ip 的个数,s 对应的是源 ip 的总数。为了方便分析和比较,取一组连续的数据包流量作为统计分析的单位,所取这组数据包的包数为 100,即 $s = 100$,通过这 100 个包的流量统计数据即可计算出相应的信息熵。

2)基于联合熵的流量异常检测。

联合熵是一种集变量之间不确定性的衡量手段。这是一种将多种变量的熵根据它们的相关性拟合后产生一个新的熵值的过程。联合熵定义如公式(2)所示:

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log_2[p(x,y)] \qquad (2)$$

文中选取了在时间 T 内采集到数据包的数量和连接数作为统计对象。设 $X_T(n)$ 为时间 T 内接收到数据的连接数,$Y_T(n)$ 为时间 T 内接收到数据包的数量,由公式(2)可知,联合熵的值需要由 $P(x,y)$ 计算得到,其中 $P(x,y) = (X_T = x, Y_T = y)$。假设 X_T 与 Y_T 满足一个关系 W_T,这个 W_T 表示:Y_T 可以看作是 X_T 按照 W_T 模型映射后不断求和得到的,公式如下:

$$Y_T(k) = \sum_{i=1}^{X_T(k)} W_T(i) \qquad (3)$$

所以可以得到:$P(x,y) = P(\sum_{i=1}^{X} W_T = y)$,由于 W_T 可以用指数分布来模拟[13],而指数分布的概率密度为:

$$E_{\lambda(Z)} = \lambda e^{-\lambda z}, z > 0 \qquad 4)$$

因此,综合公式(3)和公式(4)可知,$Y_T(k)$ 可以用伽马分布来描述。伽马分布是一种有两种参数的连续随机分布的模型,常被用来描述互联网流量的聚合关系。它有一个概率参数 λ 和一个形状参数 α。如果 α 是一个整数,这样整个分布就代表着 α 个独立并且符合指数分布随机变量的和,因为统计的连接数和数据包数都是整数,如果以 α 作为统计的数据的连接数,

那么这个伽马分布恰好就能用来描述 $Y_T(k)$。伽马分布的概率分布：

$$\Gamma_{\alpha,\lambda(Z)} = \frac{\lambda e^{-\lambda z}(\lambda z)^{\alpha-1}}{\Gamma(\alpha)}, z > 0 \quad (5)$$

所以可以得到：

$$P(x,y) = \Gamma_{x,\lambda(y)} \quad (6)$$

将公式(6)得出的值代入到公式(2)中，即可求出需要的联合熵值，从而实现分析数据的目的。

3 实验数据测试分析

3.1 测试环境

文中数据测试是在 Linux 系统下进行，测试使用到的正常流量是在网络正常情况下从互联网，如百度新浪等网站获得，异常流量通过软件 Nessus 和 hping3 模拟得到。文中分析讨论的异常流量主要来自端口扫描攻击和 DoS/DDoS 攻击。利用软件 Nessus 扫描实验主机得到端口扫描异常流量，利用 hping3 对实验主机进行攻击得到 DoS/DDoS 异常流量。实验拓扑如图 2。

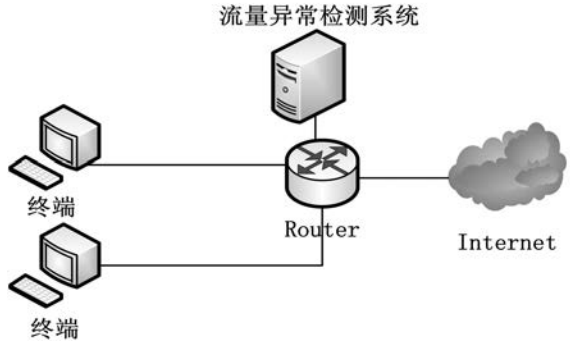

图 2 实验测试拓扑图

3.2 测试结果分析

1)基于信息熵的流量异常检测结果分析。

把 100 个包作为一组数据，连续采集了 30 组数据并计算其信息熵，同时在流量采集过程中，先采集一段时间正常流量，而后使用软件 Nessus 对实验主机进行端口扫描，从而采集到端口扫描异常流量进行流量分析。熵值曲线如图 3 所示。

从图 3 可以看出，源 ip、目的 ip、目的端口的信息熵在第 22 组数据时发生明显改变，源 ip、目的 ip 的熵值急剧下降，而目的端口的熵值急剧上升，而此时正是端口扫描攻击开始的时刻。由此可以看出，基于信息熵的流量异常检测对于端口扫描异常流量检测效果明显。

DoS 攻击的测试是通过 hping3 软件实现的，通过向目标发送大量数据包从而“淹没”主机，耗尽可用资源乃至系统崩溃，而无法对合法用户做出响应。与端口扫描一样，这里同样是以 100 个包流量信息作为一组数据分析计算熵值，但通过实验测试发现，采集到流量的信息熵值变化并不明显，即该算法对 DoS/DDoS 的检测效果不明显。下面提出了一种针对 DoS/DDoS 反应更为明显的检测方式：基于联合熵的流量异常检测。

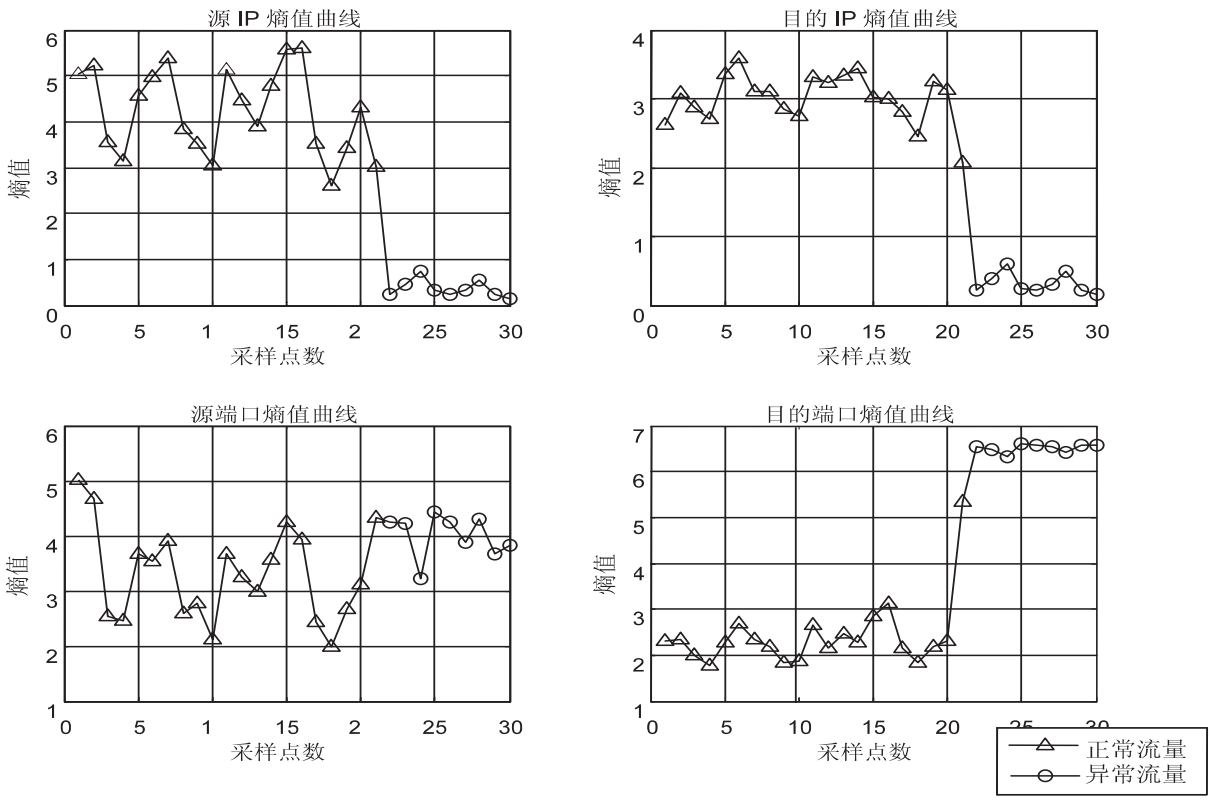

图 3 端口扫描异常流量熵值曲线

2)基于联合熵的流量异常检测。

文中在计算联合熵时将 T 选为 30s,以 30 秒采集到的数据为一组,连续采集了 7 组数据进行联合熵的分析与计算。DoS 异常流量由 hping3 攻击实验主机得到,采集过程中,先采集一段时间异常流量,而后取消攻击采集正常流量。联合熵值曲线如图 4 所示。

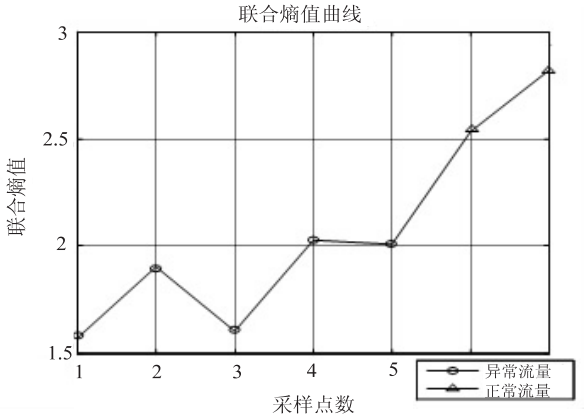

图 4　DoS 攻击联合熵值曲线

其中前 5 组为 DoS 异常流量,后两组为正常流量,可以看出,正常流量的联合熵明显大于异常流量。当异常攻击消失时,熵值明显变大。这个与基于信息熵的异常检测结果比较可以看出,基于联合熵的检测算法得到的异常与正常的对比度明显要好。这也就可以说明,在 DoS 的攻击检测下,基于联合熵的检测效果要明显好于基于信息熵的异常检测。

DDoS[14](分布式拒绝服务攻击)可以看成很多“僵尸”主机同时向被攻击主机发送 DoS 攻击。因而当 DDoS 攻击发生时,被攻击主机收到的数据包数会更多,从公式(5)和公式(6)可知,随着数据包数 y 的增加,联合熵的值会减小,所以,当 DDoS 攻击发生时,随着“僵尸”主机的增加,对应的数据包增加会更明显,这就导致对于异常流量的联合熵值将会更小,这样与正常流量的联合熵值对比更加明显,换言之,就是检测效果更好。由此可以得到基于联合熵的流量异常检测算法对于检测 DoS/DDoS 异常有较好的结果。

4　结束语

通过对基于信息熵的流量异常检测与基于联合熵的流量异常检测的实验测试与分析比较,可以看出基于信息熵的流量异常检测对于端口扫描的异常流量检测效果很好,对于 DoS 的异常检测效果不明显;但基于联合熵的流量异常检测算法对于 DoS/DDoS 异常攻击有较好的检测效果。原因在于基于联合熵的异常检测是在针对 DoS/DDoS 异常的攻击特性设计的一种检测策略:根据攻击特性选取数据包数与连接数作为分析对象构建联合熵从而达到检测目的。这就说明,基于联合熵的流量异常检测比基于信息熵的流量异常检测要精确敏感,因此,在进行异常流量分析时,需要有针对性地分析攻击特性后制定检测策略,从而可以更好地实现异常检测。同时,统计异常流量的分布特性进行训练学习,有助于对异常流量进行分类辨别,可以提高检测效率。

参考文献:

[1] 杨　策,张永智,庞正社. 网络流量监测技术及性能分析[J]. 空军工程大学学报(自然科学版),2003(2):57-60.

[2] 杨　丹. 网络流量异常检测方法研究及仿真平台研制[D]. 成都:电子科技大学,2008.

[3] 邹柏贤. 网络流量异常检测与预测方法研究[D]. 北京:中国科学院,2003.

[4] Denning D E. An Intrusion-detection Model[J]. IEEE Trans. on Software Eng. ,1987,13(2):222-232.

[5] 李　建,李　杰,孙燕花. 基于聚类融合的入侵检测[J]. 计算机技术与发展,2011,21(10):250-252.

[6] 付祖芸. 信息论-基础理论与应用[M]. 北京:电子工业出版社,2001.

[7] Zhu Yingwu, Yang Jiahai, Zhang Jinxiang. Anomaly Detection Based on Traffic Information Structure[J]. Journal of Software,2010,21(10):2573-2583.

[8] 蒲天银,秦　拯. 基于 Netflow 的流量异常检测技术研究[J]. 计算机与数字工程,2009(7):115-118.

[9] 林佳涛. 基于 NetFlow 的流量监控系统的设计与实现[D]. 广州:华南理工大学,2011.

[10] Bin L, Chuang L, Jian Q, et al. A NetFlow based flow analysis and monitoring system in enterprise networks[J]. Computer Networks,2008,52(5):1074-1092.

[11] 姜　巍. 标识分离映射网络流量监测技术研究与实现[D]. 北京:北京交通大学,2011.

[12] 陈锶奇,王　娟. 基于信息熵理论的教育网异常流量发现[J]. 计算机应用研究,2010,27(4):1434-1436.

[13] Rahmani H, Sahli N, Kamoun F. Distributed denial-of-service attack detection scheme-based joint-entropy[J]. Security and Comm. Networks,2011,5(9):1049-1061.

[14] 井艳芳. DoS 攻击的研究和主机安全防御系统的设计[D]. 青岛:山东科技大学,2004.

基于熵的流量分析和异常检测技术研究与实现

作者： 崔锡鑫， 苏伟， 刘颖
作者单位： 北京交通大学 电子信息工程学院,北京100044
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(5)