

文章编号: 1001-7445(2011)增 1-0161-04

异常检测中信息熵灵敏度分析

刘 军 程 光

- (1. 东南大学 计算机科学与工程学院, 江苏 南京 211189;
2. 计算机网络和信息集成教育部重点实验室, 江苏 南京 211189)

摘要: 信息熵的概念广泛应用于流量异常检测中, 信息熵可以用于衡量流量在某个特征分布的离散程度。基于信息熵的流量异常检测, 往往需要进行阈值选择, 用于判定实际流量特征分布的信息熵与基准分布的信息熵偏差多少就将其划分为异常。本文通过信息熵灵敏度分析来为阈值的选择提供参考。

关键词: 信息熵; 异常检测; 灵敏度

中图分类号: TP393 **文献标识码:** A

Analysis of entropy sensitivity in anomalies detection

LIU Jun , CHENG Guang

- (1. School of Computer Science & Engineering, Southeast University, Nanjing, 211189;
2. The Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, 211189)

Abstract: Entropy is widely used in traffic anomaly detection, such as measurement of the discrete level of a traffic feature distribution. Entropy-based traffic anomaly detection always face the problem of the threshold selection which is used to determine whether a real traffic contains anomalies when its entropy of a traffic feature compares to the entropy of the baseline distribution of the same feature. In this paper, we provide references for threshold selection through analyzing the sensitivity of the entropy used in anomalies detection.

Key words: entropy; anomalies detection; sensitivity

现如今, 互联网的规模越来越大, 所承载的应用也越来越多元。作为骨干网中的路由节点, 流经节点的流量高速且巨大, 流量中的网络行为也包罗万象, 有正常的网络行为, 如: web 访问、ICQ 通信、下载等, 也有异常的网络行为, 异常网络行为可以包括由于软件、设备等设计缺陷或故障造成的非恶意的异常网络行为, 如: 2009 暴风影音客户端发起的大量 DNS 请求, 也包括恶意的异常网络行为, 如: DDoS 攻击、端口扫描、蠕虫等。网络异常行为往往会浪费网络资源, 降低网络的性能, 恶意的网络行为还会造成重大网络安全问题, 如: 个人信息泄露、网络服务无法提供等, 进而能引发经济、社会和政治等问题。

鉴于网络安全的重要性和紧迫性, 研究人员一直在研究如何能从复杂的网络流量中检测出异常流量。传统的入侵检测系统其检测方法大体可以分为两种, 滥用检测和异常检测。滥用检测类似专家系统, 预先将所有网络攻击行为的模式和特征描述在一个特征库, 然后将实际流量的网络行为与特征库比较, 若有匹配, 则发现异常。Snort 和 Bro 系统就采用了滥用检测方法, 滥用检测难于检测类型新颖的

收稿日期: 2011-09-06; 修订日期: 2011-10-12

基金项目: 国家 973 研究计划(2009CB320505); 国家自然科学基金项目(60973123)

通讯联系人: 程 光 (1973-), 男, 安徽黄山人, 东南大学教授, 博导; E-mail: gcheng@njnet.edu.cn。

攻击,其时间和空间复杂度跟规则的复杂度和多少有关。异常检测预先定义正常流量行为的基准模型,当流量行为偏离正常行为基准模型则发现异常。异常检测的困难在于较难建立基准模型,误报率较高。近年来,信息论越来越多地应用于异常检测中。

信息熵是信息论中的一个概念,其被广泛应用于网络流量异常检测。Wenke Lee^[2]等人介绍了信息论中一些适用于异常检测的概念,并举例这些概念如何应用于异常检测。信息熵可以用于描述审计日志的规律性,当信息熵越小,则不同种类的记录数越少,说明审计日志越有规律。条件信息熵可以通过测量审计日志记录的顺序的规律性来进行异常检测。相对熵可以测量两个数据集规律的相似性。信息增益可以衡量某个特征划分数据集的能力。信息成本可以用于衡量异常检测模型处理数据所花费的代价。Yu Gu^[1]采用最大熵估计的方法来进行异常检测。Anukool Lakhina^[3]等人提出利用流量的特征分布进行大规模网络的流量异常检测,该方法采用了抽样信息熵来衡量特征分布的离散和集中程度。Staniford^[9]将信息熵应用于端口扫描的检测。Laura Feinstein^[10]用信息熵来检测DDoS攻击。相较于用特征匹配的方法来进行异常检测,基于信息熵的方法能检验更多类型的异常,例如:Alpha Flows、DOS、Flash Crowd、Port Scan、Network Scan、Outage Events、Point to Multipoint和Worms。在许多基于信息熵的异常检测方法中,都需要将实时流量分布的信息熵与基准分布的信息熵作比较,并且规定一个阈值,当两个信息熵的偏差大于阈值时,则判定实时流量发生了异常。这就涉及到阈值选择问题,阈值决定了可以检测多大流量分布变化。

本文提出了个新的概念——信息熵灵敏度,用于衡量信息熵检测流量分布变化的能力。本文设计这样一个实验方案:首先获得真实流量,然后真实流量中混入异常流量,通过改变异常的强度来观察信息熵的变化情况,从而获知信息熵的灵敏度。

1 相关定义

熵的概念源于热力学。在热力学中熵是大量微观粒子的位置和速度的分布概率的函数,是描述系统中大量微观粒子的无序性的宏观参数,熵越大则无序性越强,称为热熵。1948年,香农将热力学中的熵引入到信息论中。香农认为信息是人们对事物不确定性的消除或减少,不确定的程度就称为信息熵。

设随机变量 X ,其所有可能的结果是 $x_1, x_2 \dots x_n$,每种结果对应的概率是 $p_1, p_2 \dots p_n$,则其不确定程度,即信息熵是:

$$H(X) = - \sum_{i=1}^n p_i \log p_i \quad 0 \leq H(X) \leq \log |X|$$

当 X 只有一种取值情况,即是绝对值,没有不确定性,其信息熵取得最小值0。当 X 在随机结果中均匀分布时,取得最大值 $\log |X|$, $|X|$ 是随机结果的数量。

一个孤立系统的熵,自发地趋于极大,随着熵的增加,有序状态逐步变为混沌状态,不可能自发的产生新的有序结构,熵的这种性质叫着熵增原理。熵增原理预示自然界朝向无序发展。

在理想的情况下,事物是朝着无序方向发展的。但是自然界的事物是自由与约束的统一体。在外在的约束下,事物并无法发展成为最为无序的状态。与此同时,事物本身又具有一定自主性和自由度,事物总是朝着在这个自由度下所能达到的最无序的状态方向发展。事物在约束下尽可能达到最无序状态的这种性质称为最大熵原理。

最大熵统计建模是以最大熵理论为基础的一种选择模型的方法,即从符合条件的分布中选出熵最大的最优分布,即最接近的事物的真实的分布。Jaynes证明:在随机事件的所有相容预测中,熵最大的预测出现的概率占绝对优势。Tribus证明,正态分布、伽玛分布、指数分布等,都是最大熵原理的特殊情况。

2 实验分析

2.1 实验数据

本次实验计算的是报文数在源IP上分布的信息熵,计算公式如下:

$$H_{rxp}(srcIP) = \left(- \sum_{i=1}^N p_i \log(p_i) \right) / \log(N)$$

其中 N 是出现的不同源 IP 的数量 p_i 是某个源 IP 对应的报文数量占总报文数量的比例。

本次实验的数据源是 NBOS(网络行为观测系统)的中间数据, NBOS 中间数据有 NBOS 预处理模块生成, NBOS 预处理模块接收来自 CERNET 南京节点边界路由器的 NetFlow 数据, 然后对 NetFlow 进行预处理, 例如: 往返流合并、给 IP 打上归属地标签等, 生成以 5 分钟为单位粒度的 NBOS 中间数据。本次实验一共采用了连续 450 个粒度的数据。然后在这些粒度的数据中混入 4 种异常程度的单点流量, 这些异常单点流量的占总报文数的比分别为 0.5%、1%、5% 和 10%。

2.2 实验结果分析

下列的四幅图中, 横坐标表示背景流量中报文数在源 IP 上分布的信息熵, 纵坐标表示混入异常单点流量后的流量特征分布信息熵与背景流量特征分布信息熵的差。其中图 1 是混入报文数占总报文数比为 0.5% 的异常单点流量, 图 2 是混入 1% 的异常单点流量, 图 3 是混入 5% 的异常单点流量, 图 4 是混入 10% 的异常单点流量。

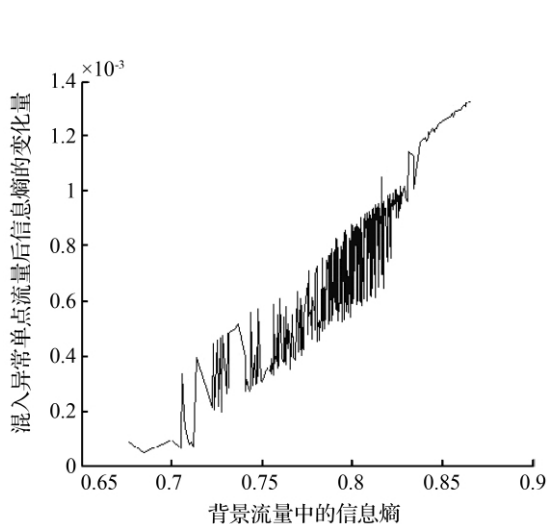


图 1 混入 0.5% 的异常单点流量

Fig. 1 mix with 0.5% abnormal traffic of a single point

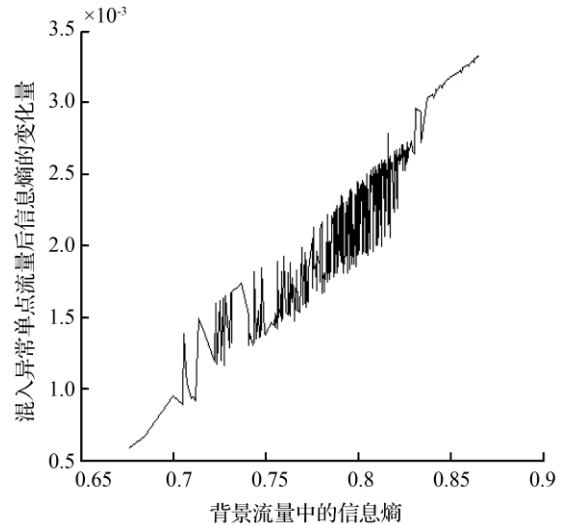


图 2 混入 1% 的异常单点流量

Fig. 2 mix with 1% abnormal traffic of a single point

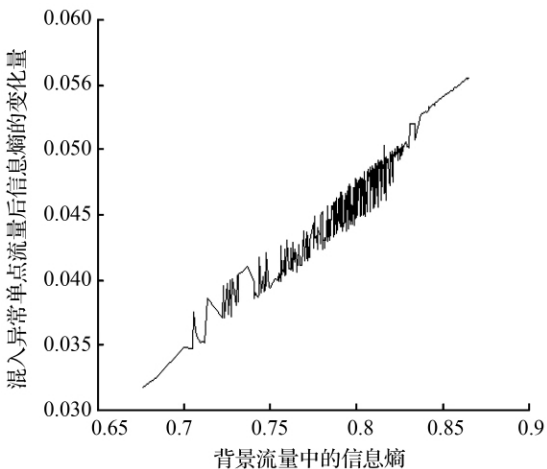


图 3 混入 5% 的异常单点流量

Fig. 3 mix with 5% abnormal traffic of a single point

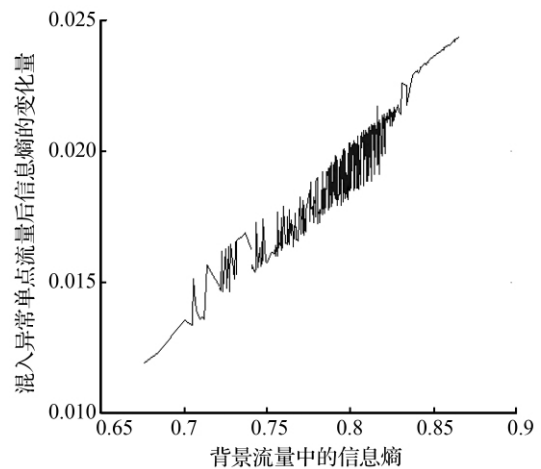


图 4 混入 10% 的异常单点流量

Fig. 4 mix with 10% abnormal traffic of a single point

从 4 幅图中我们可以看出, 混入异常单点流量之后信息熵与背景流量的信息的差随着背景流量中

的信息增大而增大,即正常流量在某个特征分布的信息熵越大,则某个异常单点的流量引起信息熵的变化越显著。

图1的纵坐标的取值范围是 $[0, 0.00014]$,图2的纵坐标的取值范围是 $[0.00005, 0.00035]$,图3的纵坐标的取值范围是 $[0.01, 0.025]$,图4的纵坐标的取值范围是 $[0.03, 0.06]$ 。这说明异常程度越大单点流量引起信息熵的变化越显著。

从上述两点结论可以看出,信息熵的灵敏度跟流量本身特征分布的信息熵的大小有关。正常流量特征分布的信息熵越大,则信息熵越灵敏。所以在阈值选择的过程中,既要考虑需要检测多小的异常,也要考虑正常流量本身特征分布的信息熵。

3 总 结

论文采用在真实流量中混入不同程度的异常单点流量并观察信息熵的变化情况的方法分析信息熵的灵敏度,得出信息熵灵敏度与真实流量本身特征分布的信息熵有关的结论,这为基于信息熵的异常检测中阈值的选择提供了重要的参考。本论文还存在一些局限性,如:没有考虑多点异常流量的情况、只考虑了使流量特征分布变集中的异常的情况,这些局限性还须通过实验进一步完善。

参考文献:

- [1] GU Y, MCCALLUM A AND TOWSLEY D. Detecting anomalies in network traffic using maximum entropy estimation [C]. Proceedings of the 5th ACM SIGCOMM conference on internet measurement. 2005: p345-350.
- [2] LEE W, XIANG D. Information-theoretic measures for anomaly detection [C]. Proceedings of the IEEE symposium on security and privacy. 2001: p130-134.
- [3] LAKHINA A, CROVELLA M AND DIOT C. Mining anomalies using traffic feature distributions [C]. Proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications. 2005: p217-228.
- [4] LAKHINA A, CROVELLA M. Diagnosing network-wide traffic anomalies [C]. Proceedings of the 2004 conference on applications, technologies, architectures, and protocols for computer communications. 2004: p219-230.
- [5] FEINSTEIN L, SCHNACKENBERG D. Statistical approaches to DDoS attack detection and response [C]. Proceedings of the DARPA information survivability conference and exposition. 2003: p303-314.
- [6] HYUN J K, JUNG C N AND JONG S J. Network traffic anomaly detection based on ratio and volume analysis [C]. International journal of computer science and network security. 2006: p190-194.
- [7] STANIFORD S, HOAGLAND J AND MCALERNEY J M. Practical automated detection of stealthy portscans [C]. Computer Security. 2002: p105-136.
- [8] FRANÇOIS J, WANG S N. BotTrack: tracking botnets using netflow and pagerank [C]. Networking 2011. 2011: p1-14

(责任编辑 唐汉民)