文章编号:1001 - 893X(2012)06 - 1018 - 04

# 利用互信息进行网络异常检测的熵特征优选*

易胜蓝

(中国西南电子技术研究所,成都 610036)

**摘 要**:首先讨论了传统流量统计分析的缺点,指出熵分析能够反映更多潜在的信息,发现传统流量统计分析不能发现的网络异常。其次,讨论了流量熵和计数熵的不同,指出两者应该配合使用,不能如现有研究中一样片面地使用其中一种。最后,用互信息法分析了两种熵的常用特征,实验发现两者分别呈现冗余状态,在剔除冗余之后检测的效率有明显提高,且不失检测准确率。

**关键词**:网络异常检测;网络流量;互信息;熵特征优选

**中图分类号**:TN915;TP393 **文献标志码**:A doi:10.3969/j.issn.1001 - 893x.2012.06.038

# Entropy Feature Selection of Network Anomaly Detection by Using Mutual Information

*YI Sheng-lan*

(Southwest China Institute of Electronic Technology, Chengdu 610036, China)

**Abstract**: Firstly, the shortcomings of traditional statistical analysis using network flow data are discussed, and it is pointed out that the entropy analysis can reflect more potential information to find out more network anomaly that can not be found by the traditional statistical analysis. Secondly, the difference between the flow entropy and count entropy is discussed and it is proposed that they should be used cooperatively and that using one of them just as existing studies is not recommended. Finally, features of the two kinds of entropy are studied bymutual information analysis. The simulations show that there is redundant in them. After redundant features are eliminated, the detection efficiency is increased significantly while the detection accuracy is maintained.

**Key words**: network anomaly detection; network traffic; mutual information; entropy feature selection

## 1 引 言

传统的流量(Traffic Volume)分析仅仅对总体流量的变化敏感,在总体流量平稳的背景下对其中各个流量特征的异常不敏感。例如,总流量不变的情况下,个别 IP 的流量变大甚至挤占其他 IP 正常使用的带宽。出现这种情况多是在白天流量高峰,本来总流量就将近满载没有上升空间。这种情况下发生分布式拒绝服务(DDoS)攻击,因为总流量没有变化不能为一般的流量异常检测方法探知。

鉴于该缺陷,研究者们[1-4]提出了基于熵理论的流量特征分析。“熵(Entropy)”这个概念最先由鲁道夫·克劳修斯(Rudolf Clausius)提出,并应用于热力学中。后来香农(Shannon)第一次将熵的概念引入到信息论中来。简单来说,熵代表一个系统的混乱程度,系统内各组成部分越混乱熵值就越大(最大值为 1),系统内部越有序则熵值就越小(最小值为 0)。就上述的例子来说,总流量不变的情况下,总流量的组成是在发生变化的。如果一个 IP 的流量变大挤占了带宽,这种情况可以看作是整个流量系统向有序发展,极限情况是总流量就等于这个 IP 的流量达

* 收稿日期:2011 - 11 - 01;修回日期:2012 - 04 - 09

到流量系统的最有序状态,这时候的熵达到最小值0。可见,当某个 IP 的流量突然变大,整体的流量熵应该是在减小,我们可以通过熵值变化的程度来判读异常的发生。

目前,在网络异常检测中使用的熵特征分为"流量熵"和"计数熵"两大类。这两种定义的熵都有人使用,但是两者之间的区别与联系还没有公开文献进行分析,导致在实际中被随意选用。但是不同实验显示这两种定义的熵其实是有不同的特点和应用场景的,不能随意选用,相互替换。其次,熵的计算复杂度大大高于传统的统计分析,有必要对现有的多个熵特征进行优化和筛选,剔除冗余的特征以提高计算效率。针对以上两个问题,本文首先从理论和实验两方面分析了这两种不同定义的熵的适用范围,进而设计了一种基于互信息的熵特征优化方法,剔除了冗余的特征。实验表明,在优选特征的条件下,熵分析的网络流量异常检测在不失准确率的情况下,检测的效率有较明显的提高。

## 2　基于互信息的熵特征优选

### 2.1　流量熵与计数熵分析

熵的定义为

$$E(x) = -\sum_{i=1}^{N} p(x_i)\lg(p(x_i))$$

其中最关键的是概率 $p$ 的计算。$p$ 被定义为 $p = \frac{v(x_i)}{V}$,其中 $v$ 是 $x_i$ 分量的度量,$V$ 是总的系统度量。例如,$V$ 代表总流量的情况下,$v(x_i)$就代表 IP 为 $x_i$ 的主机的流量。这里的流量实际上就是该分量产生的网络数据包数(Packets)。而分量可以是主机(IP)、端口(Port)、协议(Protocol)、应用(Application)等,它们被用来发现以上不同层次的网络异常。于是概率被进一步定义为(定义 1)[1]

$$p(x_i) = N_p(x_i)/n_p \tag{1}$$

式中,$N_p(x_i)$为主机、端口(源、目的)、协议或应用所占包数,$n_p$ 为总包数。

使用该定义的熵被称为流量熵。这并非唯一定义方式,网络异常研究中还有另外一种以分量出现次数为准的定义方式(定义 2)[2]:

$$p(x_i) = N_r(x_i)/n_r \tag{2}$$

式中,$N_r(x_i)$为主机、端口(源、目的)、协议或应用所占 netflow 记录数,$n_r$ 为 netflow 总记录数。

使用这种概率定义的熵被称为计数熵。仍以主机为例,这种定义下,总量 $V$ 就是计算熵的时段出现的的主机地址的总数(可重复),而分量 $v$ 则是某个主机地址在该时段重复出现的次数。

下面我们以源 IP 地址这个流量特征为例来研究两者的不同与联系。首先从定义来看,定义 1 是用每个不同 IP 所占用的报文数占总报文数的比例。可以理解为不同 IP 地址发出流量大小占总流量的比例。而定义 2 是不同 IP 地址重复的次数占总 IP 出现次数(可重复)的比例。前者着重不同 IP 在流量上表现出来的混乱程度,而后者主要是各个不同 IP 出现次数表现出来的混乱程度。

我们可以推论,定义 1 对那些流量很小但是数量众多的 IP 值不敏感,对那些小包的扫描攻击、小包的蠕虫扩散攻击识别能力较弱,优点在于能具体感知流量突然增大的 IP,对大规模 DDoS 的目标等涉及流量改变的攻击敏感。而定义 2 则对 IP 重复次数敏感,对流量信息不敏感。即某个 IP 即便异常地产生了大量流量,但是出现次数不多,根据该定义计算出来的熵值不能感知该 IP 的异常。相对地,对定义 1 不能感知的小数据包扫描、蠕虫扩散等影响多个 IP 地址的攻击。这两种定义,单独使用其中一种是很片面的,两者的结合能提供发现更多不同类型的流量异常。于是需要检查的熵特征从 4 个扩展到 8 个,即源地址流量熵、目的地址流量熵、源端口流量熵、目的端口流量熵、源地址计数熵、目的地址计数熵、源端口计数熵、目的端口计数熵。

实际还有其他关于协议和应用等的熵,但是它们一般都可以用端口的熵来代替,因为大多数的协议和应用都有对应的端口。因此,实际研究和使用中的还是地址和端口的熵值。

### 2.2　基于互信息的熵特征优选

熵分析的计算复杂度远远高于传统的简单统计的分析方法,在高速网络环境下,熵值特征获取的效率远低于传统方法,使得实际应用受到一定限制。针对这种情况,研究者一方面采用一些经典方法,例如抽样,另一方面也积极地寻找有针对性的解决方法,例如将流挖掘的相关方法进入网络熵分析中。实验发现,使用相同概率定义的熵特征具有非常大的相关性。其中一个出现异常往往连带其余 3 个同时出现异常,据此推测这些特征其实存在内部相关性。下面用信息论中的互信息理论来剔除冗余特征,特征数量的减少可以极大提高检测的效率。

互信息(Mutual Information,MI)在信息论中是作为一种衡量两个信号关联程度的尺度,后来引申为对两个随机变量间的关联程度进行统计描述。设 $MI(x,y)$ 为随机变量 $x$ 和 $y$ 的互信息,则:

$$MI(x,y)=\lg\frac{p(x,y)}{p(x)p(y)}$$

式中,$p(x)$ 和 $p(y)$ 分别是 $x$ 和 $y$ 独立出现的概率,$p(x,y)$ 是 $x$ 和 $y$ 同时出现的概率。当 $MI(x,y)>>0$ 时,表明 $x$ 和 $y$ 高度相关;当 $MI(x,y)\approx 0$ 时,表明 $x$ 和 $y$ 是弱相关,它们的同现属于偶然现象;$MI(x,y)<<0$ 时,表明 $x$ 和 $y$ 互补分布,不存在关联关系。

在本应用环境中,$x$、$y$ 分别代表同定义的 4 个流量特征中的两个。用互信息法考察它们两两间的关系。在这里重要的是要判断 4 个指标间上升下降的关系(包括其上升下降的程度)。取一段时间 4 个特征的熵流,统计其上升下降的数目,上升为 1,下降为 0。以考察上升点为例,特征 $index$ 的上升点数目被记为 $CNT^{up}(index)$。在同一时间点同时上升并且两个特征上升值之间的差异在可接受范围内,则计入同现次数,记为 $CNT^{up}(index1,index2)$。所谓可接受范围由所指定的判断因子 $\beta$ 确定:在 $t$ 时刻两个特征的熵值 $index1$ 和 $index2$,如果 $|index1-index2|<\beta$ 则被认为在可接受范围内。受考察的熵流总点数为 $size(index)$,则特征 $index1$ 和 $index2$ 独立出现的概率为

$$p^{up}(index1)=\frac{CNT^{up}(index1)}{size(index1)},$$

$$p^{up}(index2)=\frac{CNT^{up}(index2)}{size(index2)}$$

同现概率为

$$p^{up}(index1,index2)=\frac{CNT^{up}(index1,index2)}{size(index1)}$$

$$(size(index1)=size(index2))$$

则:

$$MI^{up}=\lg\left(\frac{p^{up}(index1,index2)}{p^{up}(index1)p^{up}(index2)}\right)$$

$MI^{down}$ 的获取方法类似。

当考察的变量独立的时候,两者的互信息为 0,互信息的绝对值越大表明两者越相关,完全相关时,互信息为 1。在实际应用中,一般认为大于 0.1 以上就是相关的。

## 3 实验与分析

从单位网络中心提取了大约一周的流量数据和报警记录进行实验。前者是 Netflow 格式的流量统计信息,后者是 Snort 格式的报警记录。从流量信息中提取熵特征,并做互相关检测,结果如表 1 ~ 3 所示,其中 src 代表源,dst 代表目的,ip 代表地址,port 代表端口。

**表 1 总同现概率**

Table 1 The summary probability of appearring at the same time

| Probability | srcip | srcport | dstip | dstport |
|---|---|---|---|---|
| Upprobability | 48.84% | 46.51% | 48.84% | 51.16% |
| Downprobability | 51.16% | 53.49% | 55.81% | 48.84% |

**表 2 同现概率(DOWN:下降,UP:上升)**

Table 2 The probability of appearring at the same time

| DOWN | srcip | srcport | dstip | dstport | UP | srcip | srcport | dstip | dstport |
|---|---|---|---|---|---|---|---|---|---|
| srcip | – | 48.84% | 51.16% | 48.84% | srcip | – | 44.19% | 48.84% | 48.84% |
| srcport | – | – | 48.84% | 46.51% | srcport | – | – | 44.19% | 44.19% |
| dstip | – | – | – | 48.84% | dstip | – | – | – | 48.84% |
| dstport | – | – | – | – | dstport | – | – | – | – |

**表 3 互信息值**

Table 3 The mutual information

| DOWN | srcip | srcport | dstip | dstport | UP | srcip | srcport | dstip | dstport |
|---|---|---|---|---|---|---|---|---|---|
| srcip | – | 0.251 5 | 0.253 3 | 0.291 0 | srcip | – | 0.289 0 | 0.311 2 | 0.291 0 |
| srcport | – | – | 0.213 7 | 0.250 6 | srcport | – | – | 0.289 0 | 0.268 8 |
| dstip | – | – | – | 0.253 3 | dstip | – | – | – | 0.291 0 |
| dstport | – | – | – | – | dstport | – | – | – | – |

由上可知,这 4 个流量熵的特征互信息远远大于 0,呈强相关性,只需要检测其中一个就可以代表其余。计数熵表现出同样特性,如表 4 所示(限于篇幅,略去了中间结果)。

**表 4　计数熵的互信息**

Table 4 The mutual information of count entropy

| DOWN | srcip | srcport | dstip | dstport | UP | srcip | srcport | dstip | dstport |
|---|---|---|---|---|---|---|---|---|---|
| srcip | – | 0.287 1 | 0.281 1 | 0.341 1 | srcip | – | 0.327 6 | 0.344 3 | 0.348 0 |
| srcport | – | – | 0.258 5 | 0.289 2 | srcport | – | – | 0.327 0 | 0.303 2 |
| dstip | – | – | – | 0.281 2 | dstip | – | – | – | 0.341 1 |
| dstport | – | – | – | – | dstport | – | – | – | – |

实验结果表明,在使用熵分析进行有无异常检验时,只需要进行流量熵和计数熵其中一个特征的检测即可。这里推荐用“目的地址”特征,从报警记录的相关标记来看,影响目的地址的异常较多。于是,剔除冗余后的检测特征就剩下两个:{流量目的地址熵,计数目的地址熵}。表 5 是剔除冗余特征前后的检测效率和准确率的比较,同一数据集同样的检查算法,具体算法参见文献[5]。

**表 5　特征优选前后**

Table 5 Detection efficiency and accuracy before and after feature selection

| 特征值 | 准确率/% | 提取特征用时/min | 检测用时/min |
|---|---|---|---|
| 原始 8 特征值 | 78.5 | 约 30 | 0.281 1 |
| 剔除冗余后的 2 特征值 | 72.3 | 约 7 | 0.258 5 |

由表 5 可以看出,进行特征优选后,在准确率基本保持不变的情况下,大大提高了检测的效率,这对大规模高速网络具有重要意义。

## 4　结　论

熵分析可以提供比传统流量分析具有更加精确的检测结果,但是其计算复杂度大大高于传统的简单统计分析,在高速大规模网络中这种低效果尤其不可接受。本文从保障检测的准确率和提高计算效率两方面出发,将流量熵和计数熵综合使用并用互信息优选特征,减少冗余特征。实验表明,用互信息法剔除冗余特征能够有效提高检测的效率,而不损失准确率。

## 参考文献:


[1] Nychis G, Sekar V, Andersen D G, et al. An Empirical Evaluation of Entropy – based Traffic Anomaly Detection[C]// Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement. New York, USA : ACM, 2008: 151 – 156.

[2] Lall A, Sekar V, Ogihara M, et al. Data streaming algorithms for estimating entropy of network traffic[J]. ACM Sigmetrics Performance Evaluation Review, 2006, 34(1): 145 – 156.

[3] Wagner A, Plattner B. Entropy Based Worm and Anomaly Detection in Fast IP Networks[C]//Proceedings of the 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise. Washington, DC, USA: IEEE, 2005: 145 – 156.

[4] 王海龙,杨岳湘.基于信息熵的大规模网络流量异常检测[J].计算机工程,2007,33(18):130 – 133.
WANG Hai – long, YANG Yue – xiang. Network – wide Traffic Anomaly Detection Based on Entropy[J]. Computer Engineering, 2007, 33(18): 130 – 133. (in Chinese)

[5] 王娟,靳京,钱伟中,等.基于小波分解的群落流量异常检测[J].电子测量与仪器学报,2010,24(4):365 – 370.
WANG Juan, JIN Jing, QIAN Wei – zhong, et al. Community Traffic Anomaly Detection Using Wavelet Analysis[J]. Journal of Electronic Measurement and Instrument, 2010, 24(4): 365 – 370. (in Chinese)


## 作者简介:


**易胜蓝**(1981—),女,湖南常宁人,2003 年获工学学士学位,现为工程师,主要从事航空通信领域的研究工作。

YI Sheng – lan was born in Changning, Hunan Province, in 1981. She received the B.S. degree in 2003. She is now an engineer. Her research concerns aviation communication.

Email: sly_lan@163.com