ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.10, October 2010, pp.2573−2583
doi: 10.3724/SP.J.1001.2010.03698

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于流量信息结构的异常检测*

朱应武[1,2], 杨家海[1,2+], 张金祥[1,2]

[1](清华大学 信息网络工程研究中心,北京 100084)

[2](清华信息技术国家实验室(筹),北京 100084)

Anomaly Detection Based on Traffic Information Structure

ZHU Ying-Wu[1,2], YANG Jia-Hai[1,2+], ZHANG Jin-Xiang[1,2]

[1](Network Research Center, Tsinghua University, Beijing 100084, China)

[2](Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, China)

+ Corresponding author: E-mail: yang@cernet.edu.cn

Zhu YW, Yang JH, Zhang JX. Anomaly detection based on traffic information structure. *Journal of Software*, 2010,21(10):2573−2583. http://www.jos.org.cn/1000-9825/3698.htm

Abstract: Due to the fact that the nature of network traffic is not fully and understood, large-scale, high-speed network traffic anomaly detection in an idea is a difficult problem to solve. According to the analysis of the network traffic structure and traffic information structure, it is found that in a certain range, the IP address and port distributions exhibit heavy tail and self-similar characteristics. The normal network traffic has a relatively stable structure. This structure corresponds to a more stable value of information entropy. Abnormal traffic and sample traffic of information entropy fluctuates by using the normal traffic as the center, and forms the structure of spatial information of IP, port, and IP number of active dimensions. Based on this discovery, the paper proposes a novel traffic classification algorithm, based on support vector machine (SVM) method, that transforms the traffic anomaly detection issue to a SVM-based classification decision issue. The experimental results not only evaluate its accuracy and efficiency, but also show its ability to detect on sampled traffic, which is very important for the traffic data reduction and efficient anomaly detection of high speed networks.

Key words: anomaly detection; network traffic structure; traffic information structure; anomalous traffic; sampling

摘 要: 由于人们对网络流量规律的认识还不够深入,大型高速网络流量的异常检测仍然是目前测量领域研究的一个难点问题.通过对网络流量结构和流量信息结构的研究发现,在一定范围内,正常网络流量的IP、端口等具有重尾分布和自相似特性等较为稳定的流量结构,这种结构对应的信息熵值较为稳定.异常流量和抽样流量的信息熵值以正常流量信息熵值为中心波动,构成以IP、端口和活跃IP数量为维度的空间信息结构.据此对流量进行建模,提出了基于流量信息结构的支持向量机(support vector machine,简称 SVM)的二值分类算法,其核心是将流量异常检测

* Supported by the National Basic Research Program of China under Grant No.2009CB320505 (国家重点基础研究发展计划(973)); the National High-Tech Research and Development Plan of China under Grant Nos.2007AA01Z2A2, 2009AA01Z205 (国家高技术研究发展计划(863)); the National Science and Technology Supporting Plan of China under Grant No.2008BAH37B05 (国家科技支撑计划)

Received 2009-03-30; Revised 2009-05-21; Accepted 2009-07-09

转化为基于 SVM 的分类决策问题.实验结果表明,该算法具有很高的检测效率,还初步验证了该算法的抽样检测能力.因此,将该算法应用到大型高速骨干网络具有实际意义.

关键词: 异常检测;网络流量结构;流量信息结构;异常流量;抽样

中图法分类号: TP393 **文献标识码:** A

网络流量经常会出现偏离正常范围的异常流量,主要是由蠕虫传播、DOS 攻击、DDOS 攻击、僵尸网络等恶意网络攻击行为以及网络配置失误、偶发性线路中断等引起[1].这些异常往往会导致整个网络服务质量急剧下降,使受害端主机、网络直接瘫痪.因此,如何在大规模网络环境下进行网络异常检测并及时提供预警信息,对保障网络正常运行具有重要意义.

一般来说,检测精度、实时性、全面性和新异常行为识别能力是评价异常检测系统的四大关键指标,由于网络流量在诸多层面表现出高可变性,异常流量覆盖网络攻击、网络配置失误和用户操作异常等多个范围,网络攻击行为不断出现新的方式和变种,背景噪声流量往往掩盖异常行为,使得系统的设计和实现变得十分困难.随着网络带宽的不断增加,异常检测受到测量技术、计算资源的限制,抽样检测成为大型高速网络流量异常检测的必然选择.但是,抽样不可避免地会丢失流量信息,这进一步增大了异常检测研究的难度.

本文针对检测精度、全面性和新异常行为识别能力这三大关键指标开展研究,并研究在抽样情况下如何保证检测精度.研究的核心问题是如何实现流量建模,通过挖掘流量特征,本文在流量建模方面取得了突破性进展.主要在数据采集方式上,从传统的以固定时间段的流量分片转变为按时间顺序以固定数量的数据包进行分片,发现以此作为单位流量,每个单位流量包头五元组〈协议,源地址,目的地址,源端口,目的端口〉中的源地址、目的地址、源端口、目的端口出现次数的分布符合重尾分布.比如在单位流量的源地址出现次数方面,少部分源地址出现的次数较多,大部分源地址出现的次数很少,很多源地址只出现 1 次.同一网络的一定范围内,计算正常流量中单位流量五元组的样本熵,熵值都大致相同.如果出现异常,则异常流量将改变流量的信息分布结构,熵值也将与正常情况下的熵值出现很大差别.不同类型、不同比例的异常流量,熵值也有很大不同,比如说 DDOS 攻击的目的地址熵值较小,源地址熵值较大,网络扫描探测的目的地址熵值较大,源地址熵值较小.由于正常流量具有较为稳定的信息结构,异常流量将破坏这种信息分布结构,从而可以从一个新的角度完成流量建模过程.

第 2 个重要问题是检测算法,由于各种流量都对应于不同的五元组熵值,检测异常可以转化为单位流量熵值的分类.本文提出了支持向量机(support vector machine,简称 SVM)的二值分类检测算法,算法使用核函数对五元组熵值进行核变换,产生高维线性可分数据,通过 SVM 对数据进行分类,可以达到很高的检测精度.为了解决发现新的异常流量的问题,本文提出的算法对 SVM 训练(检测)模型进行了二值分类,每一个训练模型只训练一种流量,符合该流量特征判断为 1,不符合的判断为 0.检测时,五元组熵值数据逐次通过多个训练(检测)模型,若被判断为 1,则停止检测,触发报警;若最终被判断为 0,则出现了新的异常行为.

还有一个重要问题是抽样检测.实验对流量数据的计算发现,等比例抽样不改变单位流量的自相似和重尾分布特征.抽样后仍然取单位流量五元组计算熵值,正常流量与异常流量之间的熵值差别仍然很明显.比如说相对于正常流量的五元组熵值,DDOS 攻击的目的地址熵值小,源地址熵值大.由于抽样流量的这些特性,可以根据计算能力和网络带宽自定义抽样比例.

本文第 1 节介绍异常检测的相关工作和国内外研究的一些进展情况.第 2 节讨论单位流量的统计特征和规律.第 3 节介绍异常检测算法的主要思想.第 4 节介绍实验结果.第 5 节与相关工作进行比较.

1 相关研究工作

异常检测研究分类方法较多,按检测建模方法可以分为基于误用检测(misuse-based)和基于异常检测(anomaly-based)两类;按数据采集方式可以分为抽样检测和非抽样检测两类;按检测节点数量可以分为分布式异常检测和单节点异常检测两类;按数据处理方式可以分为初级流量特征统计的异常检测和挖掘数据特征规律的异常检测两类.

误用检测通过建立攻击样本,通过描述每一种攻击的特殊模式来检测.该方法的查准率很高,并且可提供详细的攻击类型和说明,是目前入侵检测商用产品中使用的主要方法,但是需要维护一个代价较高的攻击模式库,并且存在只能检测已知攻击的弱点,一旦攻击者修改攻击特征模式来隐藏自己的行为,这种检测方法就显得无能为力.异常检测通过建立流量的正常行为模型来判断网络是否出现异常,基于流量异常的检测方法有很多种,较常用的有基于阈值、基于统计、基于小波、基于马尔可夫等随机过程模型的方法和一些基于机器学习、数据挖掘和神经网络等检测方法.

抽样检测[2−4]是大规模高速网络异常检测的必然选择,按研究重点可以分为抽样技术的研究和抽样对异常检测影响的研究.抽样技术主要分为 3 类:第 1 类是集中式抽样测量技术,抽样测量算法随机生成一个抽样事件,如以确定的计时器或计数器溢出作为激发抽样的事件,系统在报文到达之前就已经决定其是否被抽样.这种抽样方法需要时刻生成抽样事件;第 2 类是分布式抽样测量技术[2],抽样事件事先确定,在报文到达之前不能确定其是否被抽样,只有当报文到达以后根据报文内容才能决定抽样与否;第 3 类是等比例抽样,测量数据按照固定比例采集,抽样方法简单、易行.相关研究已经证明,抽样不会对异常检测产生重大影响[3],因此这类方法具有很强的实用性.由于抽样不可避免地会损失流量信息,抽样对异常检测的影响也是当前的研究重点.Brauckhoff[3]研究发现,抽样不会影响异常数据包在流量中的比例,但却会丢失数据流的信息.但其又进一步发现,即使在高比例抽样的情况下,熵仍然可以反映流量结构的变化,可以最大程度地降低抽样对异常检测的影响.

分布式异常检测研究[1,5−8]认为,分布式网络节点的流量变化具有很强的相关性,可以挖掘一些在单节点不能发现的网络异常行为.Chhabra[7]通过比较交叉链路的流量来反映流量在多个层次的联系,使用非参数估计和多重假设检验方法对这种联系进行分析,能够有效识别异常行为.

基于初级流量特征统计的异常检测[9],统计流量在包数量、流量比特等方面的比较初级的基本信息,通常采用设置阈值的方式来判断是否出现异常.挖掘数据特征规律的异常检测,通常在流量基本特征的基础上,采用熵、神经网络等方法进一步提炼流量特征规律,在此基础上进行异常检测,并采用自动分类和无监督学习技术来识别新的异常行为,使建立全面、通用的流量异常检测系统成为可能.Lakhina[1]用主元素分析方法将一组流量测量矩阵分离成正常和异常两个子空间,当异常子空间的值超过一个统计极值时就触发异常.该方法检测效果好,但却难以应用在高速网络进行在线检测;Ahmed[5]采用递归最小二乘法,并使用核函数产生不确定性变化进行在线检测,该方法检测效果与文献[1]相当,但其建立流量模型的过程较为复杂.

当前,许多重要的异常检测成果建立在流量特征规律的发现和研究的基础上,有的从结构上把流量分为具有周期性规律的正常流量、背景噪声和异常流量[10],有的从行为模式上把流量分为社会层、功能层和应用层 3 个层次[11],有的从时间上揭示工作日和休息日、白天和黑夜与流量的相关性[12],这些研究对异常检测研究具有很大的启发意义.

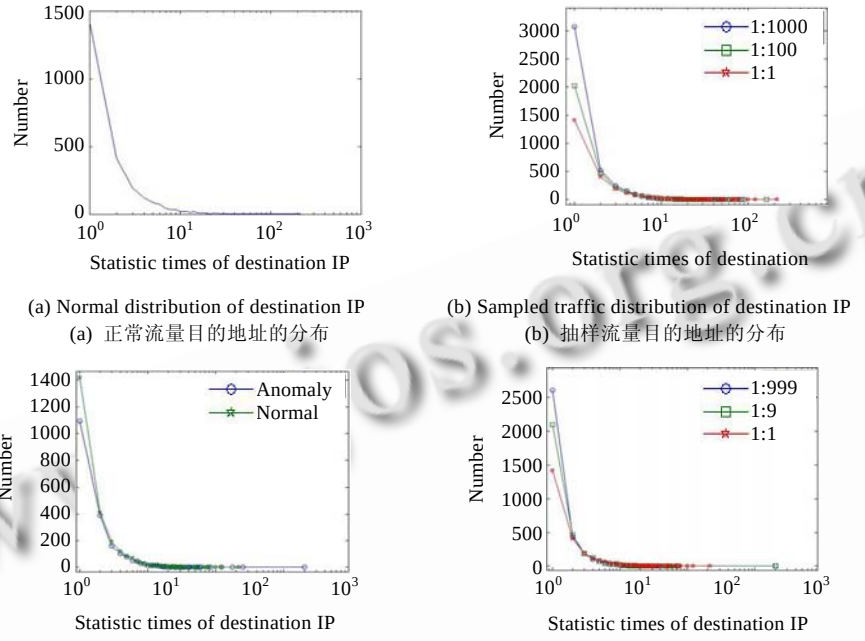
2 基于熵的流量特征分析

进行流量特征分析是异常检测和分类的强有力的工具,异常一般会显式地改变流量中 IP 地址或者端口的分布[1].很多研究人员发现,网络流量具有自相似、长相关和重尾分布等分布特征,这些发现对网络流量工程、网络建模和异常检测具有指导意义.本文研究发现,网络流量在 IP 地址和端口的分布上存在较强的重尾分布和自相似特性,不同比例异常流量和抽样流量均对 IP、端口的重尾分布产生影响.通过本文的实验可以发现,在微观层面,网络流量大部分 IP 地址的出现符合幂律分布.因此,幂律分布不仅仅涌现在网络的静态结构(网络拓扑、Web 页面链接)方面,也涌现在网络的动态流量结构方面.由于流量在时间上具有很强的突发性,单位时间的流量数据变化非常大,因此,本文的研究不以单位时间作为流量统计单位.为了更好地揭示流量的信息结构特征,下面定义几个本文常使用的概念.

- 单位流量:本文涉及到的单位流量均以一组连续的网络流量数据包作为流量的统计单位,每组数量为 10 000 个数据包,提取包头五元组〈协议,源地址,目的地址,源端口,目的端口〉,分别统计其分布规律.
- 正常流量:网络流量中的一组数据包头,经过手工和技术手段鉴别为正常的流量.

- 异常流量:按照异常流量包头五元组的分布规律,手工构造的包头五元组,并按比例注入正常流量中.
- 抽样流量:对原流量按一定比例进行等比例抽样后的网络流量,本文以等比例进行抽样.如 10%抽样表示,在原流量中,每 10 个包中提取 1 个包.

本文先对目的地址的变化规律进行研究,发现在大规模高速网络,流量在宏观上的流量变化较大,但在微观层面,单位流量具有较为稳定的分布结构,源地址、目的端口和源端口均有相同规律.图 1 中横坐标表示在 10 000 个数据包的单位流量中,某一个目的地址出现的次数,纵坐标表示出现该次数的目标地址的数量(纵坐标代表某个目的地址出现的次数,横坐标代表单位流量中出现某次数的目的地址的数量,统计量均为 10 000).



(a) Normal distribution of destination IP (b) Sampled traffic distribution of destination IP

(a) 正常流量目的地址的分布 (b) 抽样流量目的地址的分布

(c) Abnormal distribution of destination IP (d) Abnormal sampled distribution of destination IP

(c) 异常流量目的地址的分布 (d) 异常抽样流量目的地址的分布

Fig.1 Traffic feature distribution

图 1 流量的统计特征

(1) 如图 1(a)所示,单位流量中,有多达 1 400~1 500 个目的地址只出现 1 次,只有很少的目的地址出现次数超过 100 次.例如在单位流量中,常用的 Web 服务网站 IP 在数量上较少,但是每个 IP 出现的次数较多,这些分布符合重尾分布特征.源地址、源端口和目的端口也存在类似的分布规律.

(2) 如图 1(b)所示,对流量进行 1:100 和 1:1000 等比例抽样后,相对于原始流量,抽样流量具有头重尾短的重尾分布特征,如进行抽样 1:1000 后,在 10 000 个数据包中,超过 3 000 个目的地址只出现 1 次,没有出现次数超过 100 的目的地址.

(3) 相对于正常流量,异常流量具有头轻尾长、头重尾短等重尾分布特征.图 1(c)所示是在正常流量中注入 20%的 DDOS 攻击流量后,目的地址分布的变化情况.相对于正常流量,这种异常流量中只出现 1 次的目的地址的数量有所减少;但是在尾部,有目的地址出现次数等于 2 000,这个地址就是被 DDOS 攻击的目的地址,这种分布具有头轻尾长的特征.在源地址方面,异常流量的分布存在头重尾短的特征.

(4) 对异常流量进行抽样后,存在与正常流量抽样相同的变化规律.但是如图 1(d)所示,在对含有 20%的异常流量抽样后的数据中,出现次数最高的 IP 地址是被攻击的目的地址,其出现的相对频率不受抽样影响.当流量中异常流量比例大于抽样比例时,抽样具有弱化背景流量某时刻突发流量而突出异常流量的作用,从而使抽样

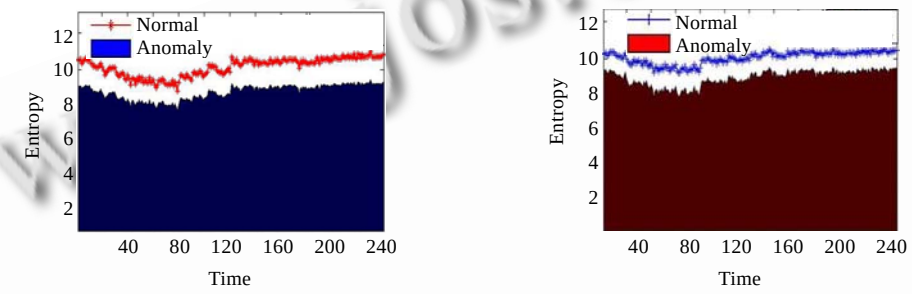
具有更高的检测精度和效率成为可能.

(5) 各种流量 IP 地址和端口分布基本服从 Hurst 参数≈0.6 的自相似性质,一定范围的异常流量和抽样流量不改变各种单位流量之间的自相似性质,正常流量的不同单位的流量之间也存在较高的自相似特征,这表明流量具有分形特征.

(6) 由于流量特征是一组五维的特征观测值,直接使用样本的观测值来区分异常比较困难,可以采用样本熵对流量信息进行度量,样本熵定义为

$$H(x)=-\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\log_2\left(\frac{n_i}{S}\right) \tag{1}$$

这里, $S=\sum_{i=1}^{N}n_i$ 是样本的全部观测值,n_i 表示某单位流量中某一协议、IP、端口出现的次数,样本熵的值位于区间 $(0,\log_2 N)$,样本全取一个值时熵取得最小值,样本全取不同值时熵取得最大值,如当 $n_1=n_2=n_3=\ldots=n_N$.正常情况下,单位流量的五元组熵较为稳定.不同比例和各种类别的异常流量在熵值上变化明显.图 2 揭示了正常流量与异常流量在熵值上的区别,其中的异常流量为根据文献[1]手工构造异常流量的方法构造的端口扫描流量,异常流量比例占 20%.由于异常流量改变了网络流量的微观结构和重尾分布特征,必然使得单位流量的平均信息熵值出现大幅变化,因此,对各种流量进行异常检测可以转化为对熵值的分类.对图 2 的数据分析还可以发现,由于网络规模和时间段的不同,影响熵值出现波动的主要原因是单位流量的源地址数量.在源地址数量一致的情况下,正常单位流量的熵值保持稳定.



(a) Entropy distribution of destination IP
(a) 目的地址的熵值分布

(b) Entropy distribution of destination port
(b) 目的端口的熵值分布

Fig.2 Feature distributions of the normal traffic and abnormal traffic induced by a port scan anomaly

图 2 端口扫描产生的异常流量与正常流量熵值的比较

3 基于熵的 SVM 的流量异常检测算法

上节分析到,IP、端口的熵值有较为稳定的常数,熵值的变化对应不同的流量类型.因此,可以使用单位流量的熵值建立正常流量和异常流量的信息模型,能够极大地简化建立网络流量模型的过程.由于低维的流量熵值模型具有线性不可分特性,本文采用核函数将其映射成高维特征空间中的线性可分数据,从而将其转化为 SVM 的分类决策问题.最后,衡量样本集、最大分类间隔和误分次数这 3 个方面的因素,取样本训练准确率最高的样本集,建立最优检测模型,从而可以运用 SVM 强大的分类预测能力对流量进行异常检测.

3.1 核函数

在建立流量信息模型后,需要将其转化为可以直接识别异常的检测模型,这一过程我们使用傅立叶核函数方法进行转换.核函数方法通过一个特征映射可以将输入空间(低维的)中的线性不可分数据映射成高维特征空间中的线性可分数据,这样就可以对高维线性可分数据使用 SVM 方法进行分类识别.核函数本质上是对应于高维空间的内积,从而与生成高维空间的特征映射一一对应.核函数方法正是借用这一对应关系隐性地使用了非线性特征映射,使我们能够利用高维空间让数据变得易于处理——不可分的变成可分的,同时又回避了高维空

间带来的维数灾难——不用显式表达特征映射.

3.2 SVM方法

支撑向量机方法以统计学习理论为基础,对于小样本学习问题,表现出很强的认知能力.SVM 主要用于模式识别、函数估计和概率密度分析.作为函数估计器,SVM 不仅与最小二乘法等传统函数估计方法有异曲同工之处,而且在估计精度和适用性方面甚至胜过后者.SVM 的基本思想可由图 3 说明,在二维两类线性可分情况下,有很多可能的线性分类器可以把这组数据分割开,但是只有一个使两类的分类间隔(图 3 中的 *margin*)最大,即图中的 *H*.

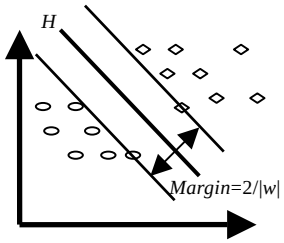


Fig.3 Optimal separating hyperplane
图 3 最优分类超平面

所谓最优分类面就是要求分类线不但能够将样本正确分开(训练错误率为 0),而且还要使分类间隔最大,这是两个相互矛盾的目标.

在进行样本分类时,每一个样本由一个向量和一个标记组成,表示为 $D_i=(x_i,y_i)$.其中,x_i 是一个高维的样本特征向量,y_i 表示分类标记.可以定义一个样本点到某个超平面的间隔 $\delta_i=y_i(wx_i+b)$.如果某个样本属于该类别,那么 $wx_i+b>0$,而 y_i 也大于 0;若不属于该类别,那么 $wx_i+b<0$,而 y_i 也小于 0,这意味着 $y_i(wx_i+b)$总是大于 0 的,而且它的值就等于$|wx_i+b|$.现在把 w 和 b 进行归一化,即用 $w/|w|$和 $b/|w|$分别代替原来的 w 和 b,那么样本点到某个超平面的间隔可以写成 $\delta_i=\frac{|g(x_i)|}{|w|}$,这就是点 x_i 到分类超平面 $g(x)=0$ 的欧氏距离,这一距离与样本的误分次数之间存在如下关系:

$$\text{误分次数}\leqslant\left(\frac{2R}{\delta}\right)^2,$$

其中,δ是样本集合到分类面的平均间隔,R 是空间中一个能够完全包含样本数据的球的半径.要降低误分次数,就要使分类间隔增大或者样本集减小,但是减少样本必然会降低训练模型的全面性和之后预测数据的准确率.另一方面,增大样本集就必然影响最大分类间隔,增大训练错误率.对流量异常检测而言,由于小异常流量对总体流量的影响并不明显,因而对特征的熵影响也不大.如图 1 所示,当异常端口扫描流量以 1:100 注入正常流量时,两者之间的熵区别很小.因此,要检测这类异常,就必须减小最大分类间隔,而这不可避免地影响到样本训练的正确率和误分次数.总的来说,样本集、最大分类间隔和误分次数是影响流量异常检测效果的 3 个因素.

3.3 基于熵的流量特征模型

第 2 节的研究发现,单位流量具有较为稳定的熵值常数,不同比例的异常流量对应不同的熵值五元组向量,抽样流量也存在类似的变化特征.因此,本文以单位流量为统计单元,建立基于 SVM 的正常和异常网络流量最优训练模型集.按照时间序列,以 IP、端口等五元组的样本熵值和源地址数量组成流量特征矩阵,即构成低维的流量特征模型,其中的每一个维度的数据变化都能反映相应的异常流量,任何一个时刻特征值的明显变化都能够表明出现异常.模型集包含训练样本数、支持的最大向量数、核函数类型和分类数等参数.最优训练模型集需要考虑样本集、最大分类间隔和误分次数 3 方面的因素.当样本训练准确率最高时,误分次数取得最小值,分类间隔最优.算法中我们选择 10 000 个流量数据报包头作为单位流量,实验的大量统计数据表明,该单位的流量具有稳定的平均信息和熵值.

3.4 算法描述

本文的算法分为 3 个阶段:第 1 阶段是对流量样本进行训练,分别得到最优训练集 ***T*** 和最优训练模型集 ***S***;第 2 阶段是检测阶段,对待检测流量按训练模型顺序进行检测,如果待检测流量能够被某一训练模型识别,则触发相应类别流量的红色警报,如果是正常流量,则不触发警报;第 3 阶段发现新的异常流量,如果在第 2 阶段中某一流量不能被当前任何一个训练模型所识别,则说明出现了新类别的异常流量,触发黄色报警,并返回第 1 阶段,将新类别流量转换为训练模型并加入集合 ***S***.算法对流量识别的核心思想是二值分类,属于某种流量的判断为

ture,反之为 false,具体流程如图 4 所示.根据第 2 节对流量熵值稳定性的分析,实际检测算法中将源地址数量加入训练集和检测集.整个算法描述如下:

Algorithm 1. Outline of SVM anomaly detection.

/*设置初始训练正确率 v_1,训练正确率 v_2,最优训练集 T,最优训练模型集 S */

/*训练阶段:分别训练 n 种异常流量,第 1 次训练正常流量,后 n–1 次训练异常流量*/

1 **for** i=1 to n **do**
2 初始化训练正确率 v_1=0,当前训练正确率 v_2=0;
3 初始化最优训练集;
4 设置训练数据集 $A_{[k][5]}$; /*$A_{[k][5]}$表示共有 k 个五维单位流量,熵值集合 B,每一个单位流量计算一个五元组熵值和源地址数组成一个六维向量*/
5 **for** m=1 to k **do**
6 计算熵值集合 $B_{[m][5]}$;
7 计算 SVM 训练的样本正确率 v_2; /*使用核函数对熵值集合 B 从六维变换到高维,再通过 SVM 进行训练,得到当前样本训练正确率 v_2 */
8 **if** $v_2 \geqslant v_1$ **then**
9 $v_1=v_2$;
10 取当前样本集为最优训练集;
11 return 最优训练模型集 $S_{[i]}$;
12 **end if**
13 **end for**
14 **end for**

/*检测阶段:设置报警标识 p,待检测的单位流量 $C_{[5]}$,计算相应的熵值向量 $D_{[5]}$ */

15 初始化报警标识 p=0;
16 取单位流量 Data:$C_{[5]}$;
17 计算熵值向量 $D_{[5]}$;

/*检测熵值向量 D,顺序通过 n 个最优训练模型集 S 进行检测*/

18 for m=1,2,…,n
19 if $D_{[5]} \in S_n$ then
20 $p=m$;
21 if p>1 then
22 触发红色警报 *Alarm*(m);
23 end if
24 return;
25 end if
26 end for

/*识别新的异常:如果 p=0,则说明出现了新的异常,触发报警并转入训练阶段*/

27 if p=0 then
28 触发黄色警报 *Alarm*(p);
29 goto 1;
30 end if

3.5 新异常流量的识别

本文的实验按照正常流量和多个异常流量的顺序建立最优训练集合,实验环境在 matlab 和 libSVM 下,每个

训练集合只训练一种流量(建立该流量在多维空间的最优分类面),若不符合该流量特征,则在 libSVM 检测时被自动识别出来,检测时按顺序检测.若当前检测集不符合任何一个训练集,则表明当前建立的训练模型与检测集在特征空间某一维度不符,表明出现了未知流量,即如图 4 所示,p=0 时所检测的流量.如本文第 1 节所述,正常流量具有稳定的信息结构,可以建立统一标准的最优训练集.因此,不符合当前最优训练集的未知流量不可能是正常流量.同时,本文对已知的异常流量已经建立训练模型,p=0 时,该未知流量也不可能是已知的异常流量,可以将其判断为新的异常流量或者已知异常流量的变种.由于绝大部分流量属于正常流量,绝大部分检测集只需要判断一次,因此本文基于二值分类的发现新异常流量的算法仍然具有非常好的平均性能.

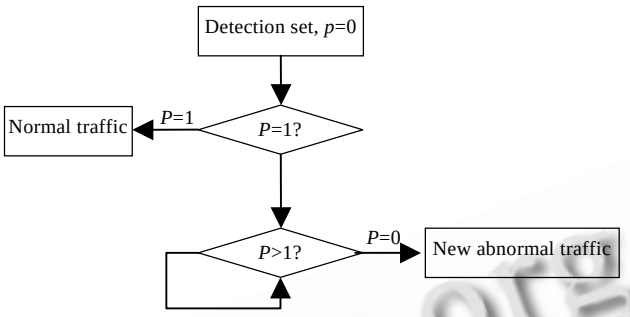


Fig.4 Process to identify new abnormal traffic

图 4 新的异常流量的识别流程

4 实验过程

4.1 实验数据

本实验数据来自清华大学校园网出口链路和 cerNet 国际出口链路采集的数据集,清华大学校园网数据采集时间为 2007 年 7 月 6 日~18 日,cerNet 国际出口链路数据采集时间为 2008 年 7 月 16 日~17 日.所有流量数据均为匿名化后的五元组〈协议,源地址,目的地址,源端口,目的端口〉.由于真实网络流量发生异常的概率较低,异常数量少,异常的规模难以刻画,因此本文所做实验对算法验证的数据由真实背景流量与异常流量构成的合成流量组成.

文献[1]对 DDOS、端口扫描、网络扫描和蠕虫病毒等异常进行了特征分析,并按照特征生成异常数据进行检测.本实验的手工异常流量按照文献[1]方法模拟生成.如在生成 DDOS 攻击流量时,按照源地址分散、目标地址集中的原则合成.一般来说,DDOS 攻击有反射式攻击、僵尸网络攻击等方式,有 Ping 攻击、TCP 半开连接等攻击,但是都满足源地址分散、目标地址集中的基本特征.这种特征与用户正常访问 Web 网站或者其他网络服务相似,但是与正常网络服务最本质的区别在于,正常网络流量的微观信息结构保持稳定,如重尾分布和 IP 地址数量.异常流量将改变这种信息结构,导致信息熵值出现很大变化.因此,可以按照手工方式合成异常流量,并按一定比例注入正常背景流量,作为实验数据的异常流量.本实验中异常流量以 10 000 个包为单位作为一个单位流量.

本文对流量微观结构和信息特征进行验证的数据量约 2T,实验第 1 分的数据量为 12 400 000 个包头五元组,实验第 2 部分的数据量约 200G,实验第 3 部分的数据量为 10 000 000 个包头五元组.

4.2 实验第1部分:对背景流量的检测

这一部分的背景流量由 12 400 000 个包头五元组组成,其中 3 000 001~12 400 000 数据包未经验证是否存在异常.异常流量为反射式 DOS 攻击流量,实验结果见表 1.

如表 1 所示,当异常流量比例分别为 1:19(5%),1:15(6.25%)和 2:23(8%)时,误检率为 0%,漏检率始终趋近于 0,但又不能达到 0.对此,我们对数据进行了手工分析发现,在实验数据集的最后一段存在 Alpha 异常流量(非正常的点对点流量),这部分流量不仅仅源地址、目的地址相同,源端口、目的端口也相同.在这一单位流量中,Alpha 流量最高能够占到总体背景流量的 25%,导致这部分流量始终被识别为异常流量.这说明,本文的算法具备在真实网络环境中的检测能力.

Table 1 Verification of cross-traffic

表 1 对背景流量是否正常的验证

Rate of abnormal traffic	Sample number of abnormal traffic	Sample number of cross-traffic	Detection rate (%)	Fall-Out rate (%)	Miss rate (%)
1:60	1 220	1 200	77.48	17.72	11.75
1:30	1 240	1 200	95.37	3.23	6.08
1:20	400	1 200	99.13	0.5	1
1:19	1 200	1 200	99.33	0	1.33
1:15	400	1 200	99.81	0	0.75
2:23	500	1 200	99.94	0	0.2

4.3 实验第2部分:对已知异常流量的验证

图 5 是算法对端口扫描、DDOS 攻击和网络扫描这 3 种异常的检测效果.实验以包含 4%异常的流量作为训练样本,经过选择不同比例样本进行训练比较.当以该比例流量作为训练样本时,样本正确率达到 99%左右,分类效果最好.该效果主要体现在如下 3 个方面:(1) 对高于 4%比例的异常的识别率基本达到 100%.如图 5 所示,当 3 种异常流量的比例达到 5%或者大于 5%时,检测率为 100%.这一效果好于文献[1]的实验效果;(2) 如果选择低于 4%的异常流量进行训练,则分类面过窄,正常流量容易被检查为异常,误报率较高;(3) 如果选择高于 4%的异常流量进行训练,则分类面过宽,对低于 4%异常流量的识别效果较差.

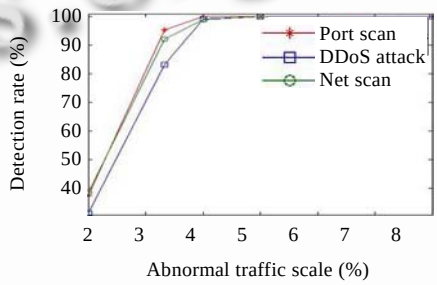


Fig.5 Detection results of the three abnormal traffic

图 5 3 种异常流量的检测效果

在图 5 所示的 3 种流量检测中,当异常流量比例从 2%~4%时,算法的检测率由 35%左右~99%,检测效果上升很快;超过 4%以后,算法的检测效果基本稳定在 100%.因此,本文将算法的检测精度定义为 4%.

本文的异常流量比例是按照数据包的数量来计算的,实际网络流量中,DOS 攻击、DDOS 攻击、端口扫描和网络扫描的包非常小,若按照流量比例计算,本文算法的检测精度更高.

4.4 实验第3部分:抽样对异常流量检测的影响

在这部分的抽样实验中,取第 1 部分前 10 000 000 个数据包作为正常流量,异常流量为反射式 DOS 攻击流量,实验结果见表 2.

如表 2 所示,在抽样率为 10%或者 1%的情况下,抽样对异常检测几乎没有影响,初步验证了按照本文的算法,抽样对异常检测效果的影响不大这样一个结论.

Table 2 The effect of sample

表 2 抽样对流量异常检测的影响

Detection rate / Rate of abnormal traffic	Sample rate 10%			Sample rate 1%		
	Detection rate (%)	Normal sample	Abnormal sample	Detection rate (%)	Normal sample	Abnormal sample
1:60	99.9	880	120	100	988	12
1:30	99.9	880	120	100	988	12
1:20	100	880	120	100	988	12

5 与相关工作的比较

由于没有一个标准的检测样本集,很多研究对异常进行手工认定,缺乏公认的比较衡量标准,也不便于在不同算法之间进行检测精度的比较.文献[1]使用了构造已知比例的异常流量来验证算法检测效果的方法,可比性强,本文的研究也使用该方法.本文基于单位流量熵值特征的异常检测研究,揭示了正常流量所包含熵值信息的稳定性和异常流量在熵值方面的变化特征,从而将异常流量的检测工作转化为熵值的分类.相对于文献[1]的主元素分析方法(PCA(principal component analysis)方法)和文献[5]的核递归最小二乘法(KRLS(kernel recursive least squares)方法),本文基于熵的 SVM 方法更简洁、高效.

(1) 在检测精度方面,本文与文献[1]均采用按比例构造已知异常流量的方法进行检测验证,文献[1]中,当 DOS 攻击的数据包在总流量数据包中的比例达到 12%以上时能够完全检测出来,当蠕虫病毒数据包在总流量数据包中的比例达到 6%以上时能够完全检测出来.本文中,当 DOS 攻击、蠕虫病毒、端口扫描和网络扫描的数据包在总流量数据中的比例达到 3.2%以上时能够完全检测出来.因此,基于单位流量熵值的 SVM 检测算法对流量数据的变化更为敏感,检测精度更高.

(2) 在检测算法的复杂性方面,文献[1]和文献[5]需要较为复杂的算法拟合正常流量模型,并需要手工设置检测极值.本文的算法只需要计算某一网络节点的单位流量熵值即可映射该网络的正常流量模型,并使用 SVM 自动地对异常流量分类.在算法性能方面,本文的算法需要的计算量为一次排序、计算熵值和 SVM 分类,约为 $O(N\times\log N)$,基本可满足在线检测的需要.

(3) 在检测全面性方面,由于各种不同的单位异常流量在五维熵值空间分布不同,本文的算法可以检测各种异常流量,并可以根据二值分类模型判断是否出现了新类别的异常行为.

(4) 相对于文献[5]的抽样方法,本文根据抽样不改变单位流量长尾分布特征和熵值稳定性的规律,抽样比例可灵活设置,更适合于根据不同带宽和计算资源选择适宜的抽样比例.

6 结 论

本文通过对各种条件下流量重尾分布、自相似等特征的分析发现,正常单位流量具有稳定的平均信息量,异常和抽样对平均信息量具有渐进性影响.以单位流量的熵值作为特征单元建立流量信息模型,使用核函数将信息模型的低维数据转换为高维线性可分数据,用 SVM 训练为二值分类检测模型,以检测模型对流量进行异常检测.通过实验发现,算法在小比例异常情况下仍然具有很高的检测率.同时还发现,在一般情况下,该算法的检测精度为 4%(即流量中异常流量的比例高于 4%时都能检测出来).另一方面,通过实验初步验证了本算法的抽样检测能力,因此将本算法应用到大型高速骨干网络具有实际意义.

References:

[1] Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions. In: Proc. of the 2005 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. Pennsylvania, 2005. 217−228.

[2] Cheng G, Gong J, Ding W. A real-time anomaly detection model based on sampling measurement in a high-speed network. Journal of Software, 2003,14(3):594−599 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/594.htm

[3] Brauckhoff D, Tellenbach B, Wagner A, May M, Lakhina A. Impact of packet sampling on anomaly detection metrics. In: Proc. of the 6th ACM SIGCOMM Conf. on Internet Measurement. Rio de Janeriro, 2006. 159−164.

[4] Mai JN, Chuah CN, Sridharan A, Ye T, Zang H. Is sampled data sufficient for anomaly detection? In: Proc. of the 6th ACM SIGCOMM Conf. on Internet Measurement. Rio de Janeriro, 2006. 165−176.

[5] Ahmed T, Coates M, Lakhina A. Multivariate online anomaly detection using kernel recursive least squares. In: Proc. of the INFOCOM, the 26th IEEE Int'l Conf. on Computer Communications. Anchorage, 2007. 625−633.

[6] Lakhina A, Crovella M, Diot C. Detecting distributed attacks using network-wide flow traffic. In: Proc. of the FloCon 2005, Analysis Workshop. 2005. http://www.cert.org/flocon/2005/presentations

[7] Chhabra P, Scott C, Kolaczyk ED, Crovella M. Distributed spatial anomaly detection. In: Proc. of the INFOCOM 2008, the 27th Conf. on Computer Communications. Phoenix, 2008. 1705−1713.

[8] Lakhina A, Crovella M, Diot C. Characterization of network-wide anomalies in traffic flows. In: Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement. Taormina, 2004. 201−206.

[9] Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. In: Proc. of the 2004 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications. Oregon, 2004. 219−230.

[10] Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark. ACM SIGCOMM Computer Communication Review, 2005,35(4):229−240.

[11] Lakhina A, Papagiannaki K, Crovella M, Diot M, Kolaczyk M, Taft N. Structural analysis of network traffic flows. In: Proc. of the Joint Int'l Conf. on Measurement and Modeling of Computer Systems. New York, 2004. 61−72.

[12] Li YQ,Yang JH, An CQ, Zhang H. Finding hierarchical heavy hitters in network measurement system. In: Proc. of the 22nd Annual ACM Symp. on Applied Computing. Seoul, 2007. 232−236.

附中文参考文献:

[2] 程光,龚俭,丁伟.基于抽样测量的高速网络实时异常检测模型.软件学报,2003,14(3):594−599. http://www.jos.org.cn/1000-9825/14/594.htm

朱应武(1978－),男,重庆人,硕士,主要研究领域为网络流量异常检测.

张金祥(1969－),男,博士,副研究员,主要研究领域为下一代互联网,分布计算,信息安全.

杨家海(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机网络,网络管理与测量,协议工程学.