一种基于非参数统计理论的网络流量异常检测方法

丁 帆,杨 越,李 军
(华中科技大学 图像识别与人工智能研究所,湖北 武汉 430074)

摘 要:提出了一种新的基于非参数高斯核函数分布的网络流量异常检测方法.与目前核函数应用于分类、神经网络、机器学习的方法和原理均不同,针对异常发生时流量出现的扰动,使用能显著反映流量形状变化的核带宽作为特征统计量,进行网络流量分析.实验结果表明,该方法能显著降低计算复杂度和误检率,提高检测率.

关键词:网络流量;异常检测;非参数统计;核函数
中图分类号:TN492 **文献标识码**:A **文章编号**:1000-7180(2011)11-0023-04

A Network Traffic Anomaly Detection Method Based on Non-parametric Statistical Theory

DING Fan, YANG Yue, LI Jun
(Institute for Pattern Recognition & Artificial Intelligence, HUST, Wuhan 430074, China)

Abstract: A new network traffic anomaly detection method based on non-parametric statistics of Gaussian kernel function distribution has been proposed in this paper. In addition, the method is different from the current theory and principle of kernel function applications, such as classification, neural network, machine learning and so on. Considering the fluctuation of the network traffic when anomaly occurs, this paper uses the bandwidth of kernel function as the feature value which can significantly reflect the change of network traffic to analyze the network traffic. Compared with other methods, experimental results show that this method can significantly reduce the computational complexity and false detection rate, also improve the detection rate.

Key words: network traffic; anomaly detection; non-parametric statistics; kernel function

1 引言

目前网络流量的检测方法[1]多采用参数统计分布模型.这对流量的平稳性要求很高,尽管有多种平稳化处理方法,但不能达到满意的效果.此外,参数分布模型的确定依赖于先验数据,经常导致参数估计及检测结果受到很大的影响.研究者很少研究流量的内在变化行为,而无攻击时,流量也会产生状态变化,状态变化的相变临界点会出现剧烈行为,而这种剧烈行为对临界点附近的流量产生长相关性影响[2],而检测过程中,该统计特征会随着流量行为的剧烈变化而变化,因而造成误检率偏高.另外,常用异常检测方法的计算复杂度较高.

针对目前统计模型存在的问题,文中提出了一种基于非参数高斯核函数分布的网络流量异常检测方法.相比参数统计模型,不仅能准确地拟合非平稳性的流量,而且能准确地估计模型参数.还能提高检测率降低误检率,可以识别出状态剧烈变化的部分正常流量.

2 基于高斯核函数的非参数统计模型

网络流量数据通常表现出非平稳性,在使用参数统计分布模型分析网络流量时,经常基于流量平稳的假设条件,要求分析数据至少宽平稳.而且该模型通常依赖于先验知识确定模型,这就要求先验数据很准确或不断更新,使其能反映其他样本数据的

收稿日期:2011-01-27;修回日期:2011-02-22
基金项目:国家自然科学基金项目(60773192)

统计性质. 已有学者研究了网络流量的性质，提出了多种参数统计模型，有 Poisson 分布、ARMA 模型、Pareto 模型等，这些模型很难准确地描述网络流量的复杂特性，而且拟合非平稳的网络流量时也存在其固有的缺点.

因此，文中选择非参数分布模型来分析网络流量，该方法要求的假定条件较少，不需要对流量做平稳化假设，由于非参数分布模型以及参数的确定都依赖于待测样本，而非先验数据，因此拟合过程更为准确，也能更准确地描述数据的性质. 考虑到非参数统计中的核函数技术能够显著地降低计算复杂度，且缺少先验知识时，高斯核函数优于其他核函数，因此这里选择高斯核函数分布作为拟合数据的分布.

目前，核函数技术广泛应用于基于机器学习的支持向量机、聚类分析、模式识别、径向基神经网络、流量分类、入侵攻击分类等领域[3]，这些应用中研究者着重于核函数的选择、构造、参数的优化、分类距离的度量等，然而目前的应用却没有考虑将核带宽作为统计特征量来解决实际问题. 文中考虑到攻击发生时，流量的形状会出现波动，而高斯核函数的带宽能准确反映流量形状的变化，因此将其作为异常检测的特征量进行分析.

设 $x,z\in X,X\subseteq R^n$，输入空间 X 通过非线性函数 ϕ 映射到空间 F，其中 $F\subseteq R^m$，n 远远小于 m，核函数满足：$K(x,z)=<\phi(x),\phi(z)>$，即核函数为非线性函数之间的内积，即 m 维高维空间的内积运算转化到了 n 维低维输入空间的计算，降低了计算复杂度，高斯核函数分布的一般形式如下：

$$K(x,x_i)=\exp\left(-\frac{\|x-x_i\|^2}{2h^2}\right)$$

其中，h 为核带宽，高斯核函数统计模型不需做任何假设和先验知识，直接从样本中获取分布信息，因而能够估计任意形状的样本序列的概率密度函数，这是参数统计模型无法比拟的优势[2]. 流量时间序列的样本 $\{x_1,x_2,\cdots,x_n\}$，概率密度函数 $p(x,h)$ 如下：

$$p(x,h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)$$

显然，概率密度函数依赖于带宽 h 的选择. 基于渐近平均方差积分误差的评价准则，高斯核函数的最佳带宽 $h=1.06*\sigma*\sqrt[5]{n}$[4]，其中 σ 为高斯核的标准差，n 为样本内元素的个数.

带宽 h 调节概率密度函数的光滑程度，h 越小，函数接近针状分布；h 越大，函数越平滑，当 h 趋于无穷大时，函数趋于均匀分布[4]. 即流量的形状处于持续光滑时，无攻击，而出现攻击时，流量的形状会出现波动，h 会出现较小、较大值的波动. 而且流量状态变化的相变临界点处，若 h 由持续光滑范围的值缓慢地变小，则由于状态变化引起的长相关性现象，h 值将出现峰顶，若 h 由持续光滑范围的值迅速变小，则可能受到攻击.

3 非参数高斯核函数分布拟合及统计分析

文中使用广泛应用于入侵检测领域的 DARPA1999 数据集. 该数据集来自麻省理工林肯实验室，共包含 5 周数据，每周包含 5 天数据. 第一周和第三周为正常流量数据，第二周数据有 18 种攻击类型. 第四周和第五周的数据包含攻击，用于测试入侵检测系统的有效性. 这里，进行周期性采样，提取单位时间内到达的数据包数量验证检测本文方法的有效性，选取单位时间为 1 秒. 为防止建模过程中，异常流量不同于正常流量的统计特性对建模过程造成偏差，文中采用 DAPRA 数据中正常流量进行分布拟合，并进行对比实验. 证实高斯核函数分布拟合流量数据的优势，对比模型的参数估计，使用适用范围广泛的最大似然估计.

文中选择不同的分布模型拟合相同的正常流量，图 1 和图 2 均为第一周第 3 天正常流量的拟合曲线，其中柱形图为用于拟合的流量数据，其他的概率密度曲线为拟合分布曲线，考虑到网络流量的重尾性，使用了 4 种性质接近的分布作为对比模型.

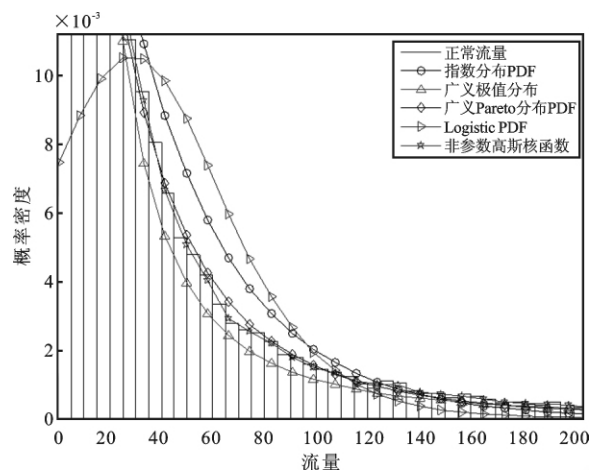

图 1 分布拟合图

显然，由图 1 可以看出几条曲线，Logistic 拟合偏差最大，广义极值分布和指数分布拟合偏差略小一些，广义 Pareto 分布和非参数高斯核函数拟合效果最好. 图 2 为拟合分布的局部图，可以看出非参数高斯核分布最优，从图像中可以看出拟合的参数统

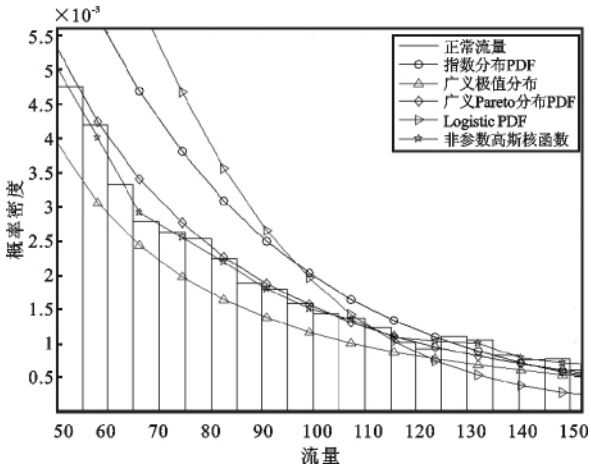

图2　分布拟合图像的局部

计模型曲线较为平滑，由于非参数高斯核分布依赖于具体样本，其形状根据样本数据不断变化，因而最为拟合数据. 另外，其他几天的分布拟合，均发现非参数高斯核函数拟合效果最优.

为了进一步确定模型的合理性，文中使用了分位数图进行分布假设检验，假设 H0：两个样本服从特定的同一连续分布；H1：两个样本服从不同的连续分布. 选择第一周第 1 天不包含攻击的流量时间序列，随机选取连续的 100 个数据作为待测数据，另在非高斯核函数分布拟合曲线选取 100 个点作为已知分布的数据，进行分布假设检验，图 3 为 QQ 图. 显然大多数点分布在虚直线上，即样本服从非参数高斯核函数分布. 图 4 为 h 取值不同时流量的分布拟合图像，显然高斯核函数的带宽 h 值影响着分布形状.

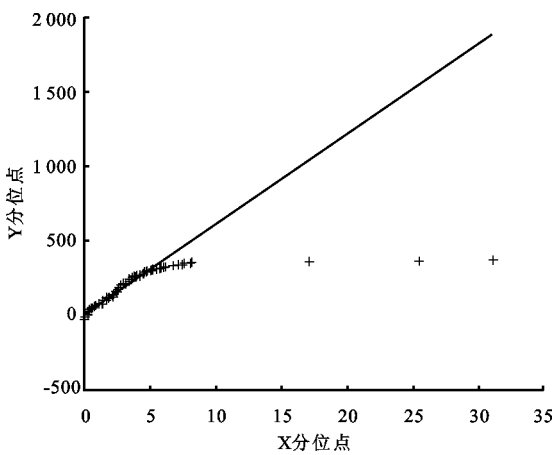

图3　QQ 图

相对参数统计模型，非参数高斯核函数有诸多优点，但是对于大样本，还是计算复杂. 因此这里对流量数据分窗，对窗内的数据建立模型，获取反映了数据变化的带宽特征量序列.

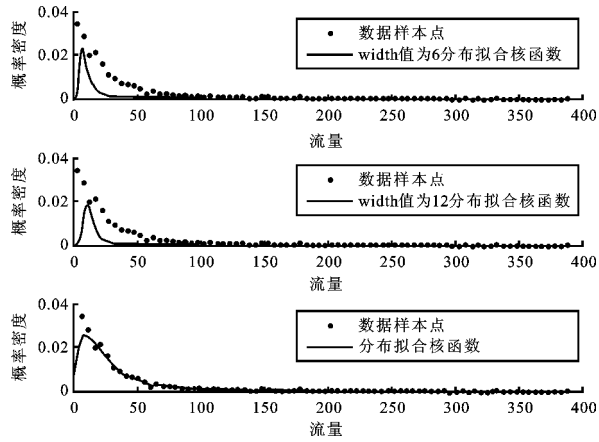

图4　带宽不同时的拟合曲线

图 5 为第一周第 1 天的流量，不包含攻击，上图为采集的 79197 个数据的流量时间图，下图为对应的高斯核带宽序列图，共 263 个窗，每个窗包含 300 个数据，显然，带宽值序列相当有规律，一天中的大部分数据的带宽值是宽平稳的，多天数据的统计均发现，当带宽特征值较大且在一定范围内波动，流量持续光滑，这时无论参数 h 变化快或慢，是否有峰值，流量均处于正常无攻击状态，而且这段光滑区域的带宽值缓慢降低之后，虽然有峰值出现，但不是攻击，显然这些实验结果与前面的分析吻合.

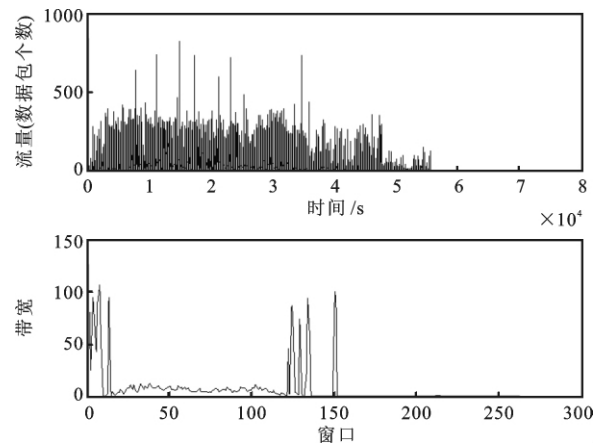

图5　W1D1 正常流量的流量与带宽值对比图

图 6 为第四周第 2 天的流量，包含攻击，上图为流量时间序列，下图为该流量的带宽序列与攻击的对比图，包含 198 个窗，每个窗包含 400 个数据，显然，椭圆形标记的为无攻击区域，攻击均发生在特征值剧烈波动的窗口，与前面分析吻合.

综合理论分析及实验验证[5-6]，带宽值的统计规律能明显反映流量正常和异常的区别. 根据带宽特征值序列，可以直观地区分宽平稳无攻击和易出现攻击区域，这大大降低了异常检测的计算复杂度，并降低了误检率.

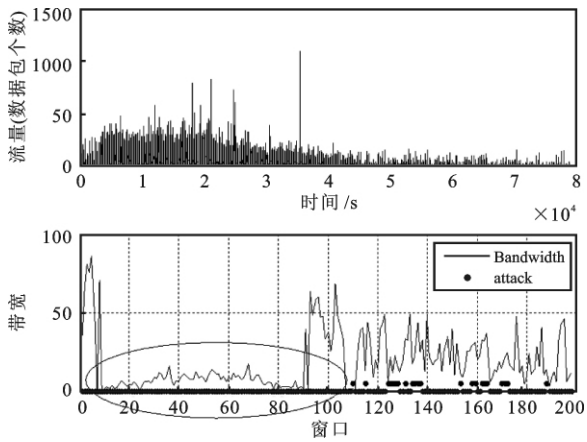

图 6 W4D2 包含攻击的流量与带宽值对比

4 结束语

文中针对参数统计模型存在平稳假设的限制条件，以及异常检测方法存在计算复杂度偏高、误检率较高的问题，提出了基于非参数高斯核函数分布模型的异常检测方法. 充分考虑到攻击出现时流量形状的变化，着重研究了描述流量形状变化的核函数带宽特征值序列，结合图像观测和检测特征值序列，极大地降低了计算复杂度，得到了良好的实验结果.

参考文献：

[1] 付雄，彭冰. 基于 Shadowing 模型的无线入侵主机物理定位研究[J]. 微电子学与计算机，2010，27(12)：4－10.

[2] 袁坚. 计算机网络中的相变和自组织临界现象[R]. 北京：清华大学，2000.

[3] 肖秀春，姜孝华，张雨浓. 一种基函数神经网络最优隐神经元数目快速确定算法[J]. 微电子学与计算机，2010，27(1)：106－109.

[4] 杨永生，张宗杰. 基于核函数和带宽的海杂波概率密度函数估计[J]. 探测与控制学报，2010，32(5)：38－41.

[5] Yang Yue, Hu Hanping, Xiong Wei, et al. A novel network traffic anomaly detection model based on superstatistics theory[J]. Journal of Networks, 2011(9): 4757－4759.

[6] Yang Yue, Hu Hanping, Xiong Wei. Network traffic anomaly detection method based on feature of catastrophe theory[J]. Chinese Physics Letters, 2010, 27(6): 2070－2074.

作者简介：

丁 帆 硕士研究生.

杨 越 博士，讲师.

李 军 博士研究生，讲师. 研究方向为网络与信息安全.

(上接第 22 页)

[2] 曹奉祥，李永明，孙义和. 一种低功耗、高线性、双正交可调谐 CMOS 上变频混频器[J]. 微电子学与计算机，2007，24(5)：166－170.

[3] Darabi H, Abidi A A. Noise in RF－CMOS mixers: a simple physical model [J]. IEEE J. Solid－State Circuits, 2000(35): 15－25.

[4] Terrovitis M T, Meyer R G. Noise in current－commutating CMOS mixers [J]. IEEE J. Solid－State Circuits, 1999(34): 772－783.

[5] Gardner W A. Introduction to random processes with applications to Signals and systems [M]. 2nd ed. New York: McGraw－Hill, 1989.

[6] Lee T H. The design of CMOS radio－frequency integrated circuits[M]. 2nd ed. New York: Cambridge University Press, 2004.

[7] Li Zhiyuan, Ma Jianguo. Compact channel noise models for deep－submicron mosfet [J]. IEEE trans. Electron devices, 2009(56): 1300－1307.

作者简介：

郭本青 男，(1977－)，博士研究生. 研究方向为射频模拟集成电路技术.

文光俊 男，(1964－)，教授，博士生导师. 研究方向为无线通信射频集成电路与系统、模数混合集成电路、计算电磁学等.