基于 K－means 聚类的网络流量异常检测

许晓东[1]， 杨燕[2]， 李刚[2]
（1. 江苏大学信息化中心，镇江 212013；2. 江苏大学计算机科学与通信工程学院，镇江 212013）

摘要： 针对网络异常检测领域存在的漏报率和误报率较高的问题，提出一种基于 K－means 聚类的网络流量异常检测方法。选择了多个不同维度上的特征；计算各维特征在滑动窗口中的局部均值偏差，以保证在实时动态变化的网络中的检测准确度；利用由 K－means 聚类算法产生的检测模型对各维特征进行综合评判，有效地降低了漏报率和误报率。在网络流量数据集上对所提方法进行了验证并和已有方法进行了对比，所提方法在精度和效率方面取得了较好的实验效果。

关键： 网络流量；异常检测；K－means；聚类
中图分类号： TP393.08 **文献标识码** A **文章编号：** 1003-8329(2013)04-0021-06

Network Traffic Anomaly Detection Based on K－means Clustering

XU Xiao-dong' YANG Yan[2]， LI Gang[2]
（1. Information Center，Jiangsu University，Zhenjiang 212013，China；2. School of Computer Science and Telecommunication Engineering，Jiangsu University，Zhenjiang 212013，China）

Abstract: In order to satisfy the demands of high true positive rate and low false positive rate in network traffic anomaly detection，a method based on K－means clustering is proposed. According to the characteristic of network traffic，several features of different dimensions are selected. In order to achieve high accuracy in dynamic networks，local deviation from mean is calculated in the sliding window. The detection model which generated by K－means clustering algorithm is utilized to fuse multi－dimensional features to decide whether the network traffic is normal，and by such fusion it achieves low missing rate and false alarm rate. The proposed method is validated and evaluated by comparing it with existed algorithms derived from some network traffic datasets. The experiments show that the proposed method can detect attacks with high accuracy and high efficiency.

Key words: network traffic； anomaly detection； K－means ；clustering

1 引 言

目前针对网络流量异常检测的研究已有很多，研究者根据异常可能引起不同特征的变化，提出了不同的检测方法，如文献 [1] 从数据流中提取服务请求，根据服务请求的三个属性：请求类型、请求长度和负载分布计算异常得分；文献 [2，3] 利用网络流量异常会引起数据包头部特征（源/目的 IP、源/目的端口等）的分布发生显著变化，引入特征熵作为异常检测的数据源。这些方法虽然采用了多个特征以提高检测率，但是没有融合多个特征进行综合评判，仅采用简单的选举法或简单的特征组合公式，而没有任何理论依据。文献 [4] 基于 D－S 理论融

* 基金项目：国家自然基金项目（61005017）。
作者简介：许晓东，教授，主研领域：网络安全，网络管理。

合多个特征对网络流量进行综合评判，有效地降低了误报率和漏报率。文献[5]提出了基于聚类的异常检测框架，首先对训练数据集进行划分，生成 n 个簇类，并根据簇类的大小标记"正常簇"或"异常簇"，然后利用已标志的簇将网络数据划分到距离最近的簇中。文献[6]在文献[5]的基础上提出一种新的方法计算类簇的半径阈值，并且根据每个类簇偏离整体的程度来区分正常簇和异常簇。

上述方法有些使用多个特征以加大检测范围，但是没有融合多个特征进行综合评判；聚类算法虽然可以处理多维数据，但是应用在网络异常检测中存在一些缺陷。针对以上问题，本文试图采用能从多个角度反映网络流量状态的特征以检测到更多的异常；为了能够融合各维特征的检测信息，降低误报率，将各维特征排列成检测向量，采用由 K - means 聚类算法产生的检测模型对各维特征进行综合判断。其中，为了使 K - means 聚类算法能够保证在实时动态变化的网络流量中检测的准确性，利用滑动窗口机制计算各维特征的局部均值偏差，使其能够体现各维特征当前的异常程度。

2 异常流量特征分析

在网络异常检测的特征选择问题上已有学者进行了深入研究，有些方法[2,3,7]重点关注数据包头部特征(源/目的IP，源/目的端口号等)；有些方法[1,8]使用了数据包应用负载中的特征进行检测，虽然利用这些特征有助于提高检测率，但由于应用负载的数据量过大，使用这些特征往往导致检测算法不能满足高速网络的检测需求。因此应该选择计算量小，能从多个不同的角度反映网络流量状态的特征。本文重点关注的维度有：源/目的IP、源/目的端口、协议类型以及TCP协议的标志位等。

相关研究[3]表明，正常情况下，在给定的时间粒度内，不同的源/目的IP地址数和不同的源/目的端口数相对稳定，并且他们之间存在一定的内在关联；当发生异常时，它们之间的关联将会被打破。文献[6]利用Hellinger距离衡量每个时间窗口内各特征的分布变化，本文对其本质进行分析，发现它衡量的是源/目的IP地址数，源/目的端口数的比例关系是否发生变化，于是本文将IP维和端口维的测度定义如下：

定义1　D_{sip}，D_{dip} 分别为单位时间内不同的源IP地址数目，目的IP地址数目，则 D_{sip}，D_{dip} 之间的比例关系为

$$H_{ip} = D_{sip}/D_{dip}$$

定义2　D_{spt}，D_{dpt} 分别为单位时间内不同的源端口号数目，目的端口号数目，则 D_{spt}，D_{dpt} 之间的比例关系为

$$H_{port} = D_{spt}/D_{dpt}$$

利用 H_{ip} 和 H_{port} 可以检测到能够引起源/目的IP地址数或源/目的端口数发生大幅度变化的异常，例如，DDoS、蠕虫、端口扫描等；但是它们没有考虑数据包的多少，因此对各特征元素个数变化不大但是会引起流量突变的异常无能为力，例如，某个主机突然向另一主机发送大量的数据包。为了弥补此缺陷，需要引入其他测度。

Moore等人通过对被攻击者发送的响应包进行分析得出：大多数攻击使用TCP包(94%以上)，其次是UDP包(2%)和ICMP包(2%)[9]，也就是说大多数情况下如果发生攻击，TCP、UDP、ICMP包中至少有一种会发生异常。而TCP，UDP和ICMP报文在网络流量中的分布具有很强的规律性。因此可以使用TCP、UDP、ICMP包的比例关系来描述网络运行情况。若发生基于UDP或ICMP协议的洪流攻击，不同协议报文的分布即会发生明显变化，但由于TCP报文一般在网络流量中占有较高比例，发生基于TCP的攻击时，3种数据包的比例关系与正常情况下基本上是相同的，因而检测不出基于TCP的攻击[10]。考虑到基于TCP的攻击通常会引起SYN报文和SYN + ACK报文的数量不匹配[1]，因此将SYN/SYN + ACK报文的对称性作为检测基于TCP攻击的测度。如上所述，协议维和TCP标志位维的检测测度定义如下：

定义3　设 P_{TCP}，P_{IP} 分别表示在单位时间内TCP报文和IP报文的统计数，则TCP报文所占比例为

$$H_{tcp} = P_{TCP}/P_{IP}$$

定义4　设 P_{SYN}，$P_{SYN+ACK}$ 分别为单位时间内SYN和SYN + ACK的报文数，则SYN和SYN + ACK对称性为

$$H_{syn} = P_{SYN}/P_{SYN+ACK}$$

3 异常检测方法

各维特征度量值的时间序列在正常情况下比较平稳，当有异常发生时会发生较大的波动，当超过一定阈值时，则认为出现异常。但是如何确定各维特征的阈值，如何根据多维特征的检测结果来判断是否真的发生异常？为此将各维特征量排列成检测向量，采用 K－means 聚类算法从大量样本数据中挖掘出正常流量的类簇，将其作为检测模型，根据距离判断检测向量是否异常。异常检测流程如图 1 所示，整个流程分为两阶段：1）训练阶段，从样本数据中提取特征，对各维度量值进行预处理并构造检测向量，利用 K－means 聚类算法对检测向量进行分类，过滤异常数据并建立正常流量模型；2）在线检测阶段，类似地构造检测向量，利用训练阶段建立的正常流量模型对检测向量进行分类，如果检测向量不属于任何类簇，则判断发生异常。

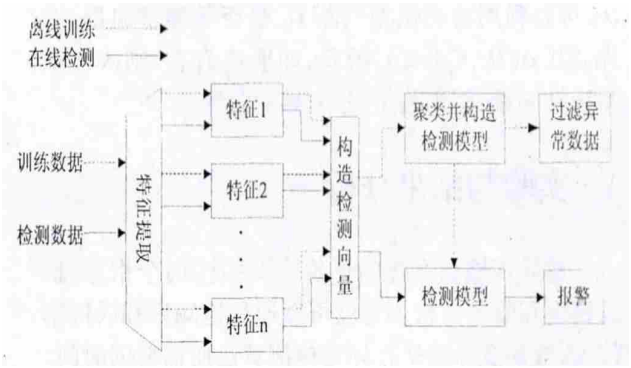


图 1　异常检测算法流程

3.1　构造检测向量

由于下一阶段采用的是聚类算法，数据的度量范围不同会对聚类结果产生影响，因此在聚类分析之前需要对数据进行标准化处理。

3.1.1　数据预处理

文中所采用的测度均是区间标度变量，对于这类变量的标准化处理方法有很多，z 分数规范化方法[6] 是最常用的一种。使用 z 分数规范化需要计算所有分类数据的均值，应用在网络流量异常检测中存在以下问题：1）网络流量是随时间依次到达并且具有无限性，因此无法计算所有数据的均值；2）网络流量随时间动态变化，各维特征的度量值在局部可能表现异常，但是在全局数据中却表现正常；或者在局部表现正常的数据在全局中表现异常。为了解决以上问题，考虑到一般情况下网络流量在一定的时间内，随时间变化较小，因此可以利用最近 a 个历史数据的均值作为当前时刻的参考值，通过计算当前度量值与均值的偏离程度衡量各维特征的异常程度。这里采用滑动窗口保存并更新历史数据。

定义 5　设滑动窗口的大小为 a，x_i 是特征 x 当前的度量值，m_x 和 σ_x 分别为特征 x 此时在滑动窗口中的均值和标准差，则特征 x 的局部均值偏差为

$$z(x) = \frac{x_i - m_x}{\sigma_x}$$

$z(x)$ 表示当前度量值偏离均值多少个标准差。由于采用了滑动窗口机制，会不断地更新历史数据，$z(x)$ 反映的是各维特征的当前异常程度，因此能够检测到局部异常。

3.1.2　剔除异常点

在网络流量正常的情况下，每隔一个单位时间，将最新的度量值加入到滑动窗口并将最旧的数据剔除，保持滑动窗口长度 a 不变。在检测到可疑点时，为了确保后续检测的精度，计算均值时会剔除可疑点。假设数组 $T[n] = \{x_1, x_2, \cdots, x_a\}$ 记录前 a 个时间窗口中特征 x 的度量值，对于当前观察值 x_j，如果 $|z(x)|$ 超过阈值 l（l 通常取值为 3），则不更新数组 T，否则用 x_j 替换 T 中最旧的数据。详细过程如下所示：

1）计算前 a 个时间窗口中特征 x 的均值 m_x 和标准方差 σ_x：

$$m_x = \frac{\sum_{i=1}^{a} x_i}{a}, \sigma_x = \sqrt{\frac{1}{a-1}\sum_{i=1}^{a}(x_i - m_x)^2}$$

2）If $\frac{|x_j - m_x|}{\sigma_x} \leqslant l$；

3）$T[n] = x_j$；

4）$n = (n+1)\%a$；

5）Else；

6）不更新数组 T；

7）End If。

3.1.3　构造检测向量

在对各维度量值进行预处理后，可以将其排列成检测向量，作为下一阶段输入的数据对象。

定义 6　设 $z(H_{ip})$，$z(H_{port})$，$z(H_{tcp})$，$z(H_{syn})$ 分别为 t 时刻 IP 维，端口维，协议维和 TCP 标志位维

测度的局部均值偏差，则可构成四维检测向量

$$D_t = < z(H_{ip}), z(H_{port}), z(H_{tcp}), z(H_{syn}) >$$

3.2 K－means 聚类算法

K－means 是一种基于划分的动态聚类算法，将其用于异常检测主要基于以下 2 个假设[5]：1）正常行为数据量要远远超过异常行为数据量；2）正常行为数据与异常行为数据之间的差异很大。第一个假设为识别正常簇与异常簇提供了依据，根据第二个假设能够很好地区分正常流量和异常流量。为了后续使用方便，下面对涉及的概念进行定义。

聚类中对象间的相异度（或相似度）通常基于对象间的距离来计算。常用的计算距离的方法包括欧几里得距离、曼哈顿距离以及闵可夫斯基距离等。文中采用欧几里得距离计算相异度。

定义 7 数据对象 i 和 j 的相异度（或距离）为

$$d(i,j) = \left(\sum_{k=1}^{n}(x_{ik} - x_{jk})^2\right)^{1/2} \quad (1)$$

其中 $i = (x_{i1}, x_{i2}, \cdots, x_{in})$ 和 $j = (x_{j1}, x_{j2}, \cdots, x_{jn})$ 是两个 n 维的数据对象，在本文中 n 取 4。

定义 8 设 $C = \{C_1, C_2, \cdots, C_k\}$ 是对训练数据集 S 的一种划分，类 C_i 的异常因子 $OF(C_i)$ 定义为 C_i 与其他类间距离的平均值：

$$OF(C_i) = \frac{\sum_{j \neq i} d(C_i, C_j)}{k-1}$$

异常因子 $OF(C_i)$ 度量了 C_i 与其他类簇间的偏离程度，$OF(C_i)$ 越大，C_i 与其他类簇间的偏离程度越大。

定义 9 设 $C = \{C_1, C_2, \cdots, C_k\}$ 是对训练数据集 S 的一种划分，类 C_i 的概要信息 $S(C_i) = \{cen_i, r_i\}$，其中 cen_i 是类 C_i 的聚类中心，r_i 是 C_i 的半径，即类 C_i 中的数据对象离质心 cen_i 最远的距离。

3.2.1 K－means 基本思想

在训练阶段，采用经典的 K－means 聚类算法对数据集 S 进行划分，其基本思想如下：

输入：n 个四维空间数据对象 $S = \{D_t \mid t = 1, 2, \cdots, n\}$，聚类中心的个数 k；

输出：k 个类簇 $C = \{C_1, C_2, \cdots, C_k\}$。

1）从 n 个对象中任意选择 k 个对象作为初始聚类中心；

2）根据每个聚类中心，计算每个对象与这些聚类中心的距离；并根据最小距离重新对相应对象进行划分；

3）重新计算每个（有变化）簇的聚类中心；

4）循环 2）到 3）直至每个簇不再发生变化为止。

3.2.2 建立检测模型

在聚类之后需要过滤掉异常数据，建立正常数据模型。其详细描述如下：

1）将集合 C 中的簇按异常因子大小升序排序，使得 $OF(C_1) \leqslant OF(C_2) \leqslant \cdots \leqslant OF(C_k)$；

2）寻找最小的 x 使其满足 $\frac{\sum_{i=1}^{x} |C_i|}{n} < y$；

3）标记 $C_1, C_2, \cdots, C_x$ 为正常簇；

4）由正常簇的概要信息构成检测模型。

这里假设训练数据集中正常数据占所有数据的比例为 y。

3.2.3 在线检测

此时，对每个时间窗口中生成的四维检测向量 D_t，可以利用检测模型判断 D_t 是否是异常向量。首先寻找 $d(D_t, C_j) \leqslant r_j$ 的 C_j；如果 C_j 存在，则认为 D_t 是正常向量，否则判定 D_t 为异常向量。

4 实验与结果分析

对异常检测系统的评价主要关注两个指标：检测精度和效率。检测精度可以通过检测率和误报率体现，效率意味着运行异常检测算法所需要的时间。本文通过与基于熵（Entropy）的算法和基于指数加权移动平均（EWMA）的算法进行对比实验验证所提方法的精度，并通过分析所提算法的时间复杂度和单步执行时间验证算法的效率。

4.1 实验数据

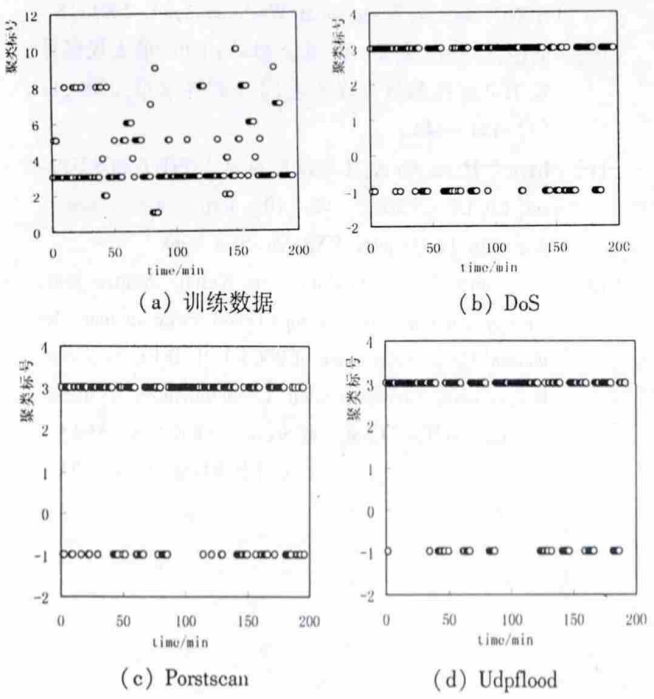
实验采用的是 NUST Traffic Dataset[11]，该数据集是由 Irfan 等人采集的校园网数据，包含 3 个数据集，每个数据集持续大约 3 个小时。在每个数据集中只施放一种攻击（portscan/DoS/Udp flood），每种攻击发生 10 次，每次持续 5 分钟，并且每种攻击具有 5 种不同的攻击速率：三种低速率攻击（{0.1，1，10} pkts/sec）和两种高速率攻击（{100，1000} pkts/sec）。其背景流量的平均速率是 3168pkts/sec，标准偏移是 1683pkts/sec。各攻击特征如表 1 所示。

表 1　攻击特征

属　性	DoS	Portscan	Udpflood
数据包个数	8611333	13844392	12310421
数据集大小	82.5MB	129MB	111MB
持续时间/s	12000	12068	11681
攻击特征	两个远程服务器遭受攻击，被攻击的端口号是: 143，22，138，137，21	固定的源 IP 地址发起的两种不同攻击: 第一次扫描端口 80，第二次扫描端口 135	两个远程服务器遭受攻击，被攻击的端口号是: 22，80，135，143

4.2　检测性能

将含有 DoS 攻击的数据作为训练数据，以 1min 时间间隔统计各维特征量。假设训练数据被分成 n 个时间间隔，则可以得到 n* 4 的矩阵(n 个检测向量)，用 K－means 聚类算法对这 n 个检测向量进行分类。当 k 取值为 10 时，其结果如图 2 所示。标号为 3 的簇包含的数据对象最多，其他簇只有极少的数据对象，根据标类算法选取标号为 3 的簇作为检测模型。然后利用检测模型分别对三个数据集进行在线分类。检测结果如图 2 的(b) 、(c) 、(d) 所示，图中标号为 3 的簇代表的是正常簇，－1 表示检测到的异常。

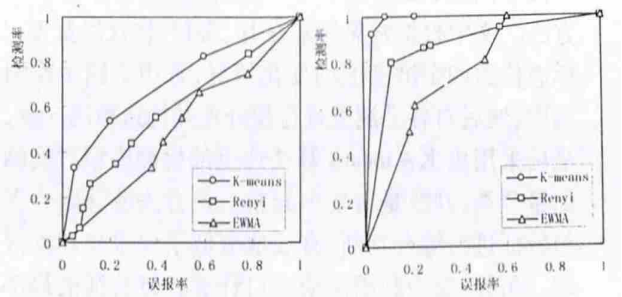


(a) 训练数据　(b) DoS

(c) Porstscan　(d) Udpflood

图 2　聚类结果

为了验证所提算法的准确性，采用近年来受关注较多的基于熵的检测方法[12] 以及基于 EWMA 流量预测模型的检测方法进行对比实验，实验结果如图 3 所示。图 3 是 3 个方法分别对低速率攻击和高速率攻击的检测结果。从图 3 中可以看到，不论是对低速率攻击还是高速率攻击，所提算法在大部分情况下相对于其他 2 种方法具有更高的精度，而 EWMA 方法的检测精度最差。主要原因分析如下: 1) 流量大小在时间序列上存在较大的波动，因此基于 EWMA 的算法的误报率较高，而本文所采用的特征序列在无异常时相对平稳，并且采用 K－means 算法综合各维特征的检测信息，因此误报率更低; 2) 在计算局部均值偏差时剔除了异常数据，使其能够检测到多个连续的异常，因而检测率更高。另外，在低速率攻击中，三种方法的检测率相对较低的原因是低速率攻击的网络行为已经非常接近正常流量的网络行为，并未对整体流量产生较大的偏移。通过调整阈值，虽然可以提高检测率，但是误报率会相应上升。



(a) 低速率攻击　(b) 高速率攻击

图 3　网络流量数据集 ROC 曲线

4.3　实验效率分析

通过对算法的仔细分析发现进行异常检测所需要的时间主要消耗在特征量的统计上，在流量较大的网络上，统计特征量对算法的效率影响较大。在算法实现时，利用哈希表记录已出现的特征元素，并用链地址法解决冲突问题。K－means 聚类算法在训练阶段的时间复杂度是 $O(nkt)$，其中 n 是对象个数，k 为聚类个数，t 为迭代次数; 建立检测模型阶段需要计算各个类簇的异常因子并对其进行升序排序，时间复杂度为 $O(m \cdot k^2)$，其中 m 是检测向量的维度; 在线检测阶段的时间复杂度为 $O(m \cdot x)$，其中 x 是正常簇的个数。

在配置为 1.9GHz 的 CPU、1GB 内存的计算机

上，对以上实验数据集应用所提方法，单步执行时间如图 4 所示。对其进行分析，发现在正常情况下所提方法的单步执行时间不超过 1s，在发生大规模网络攻击时单步执行时间不超过 20s，具有较好的实时性。

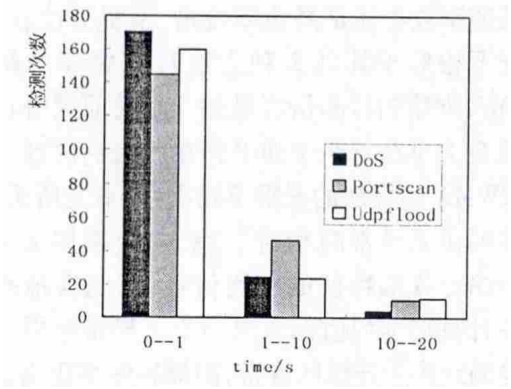


图 4　算法在各数据集上的运行时间

5　结束语

针对网络流量检测领域检测率较低，误报率较高两个问题，提出基于 K－means 聚类的异常检测方法。文中根据常见异常在 IP、端口、协议以及 TCP 标志位这四个维度上的变化情况，采用了四个检测测度，然后对各维测度进行预处理并构造检测向量，最后采用由 K－means 算法产生的检测模型对检测向量分类，判断是否发生异常。该方法能够融合多维特征进行综合判断，有效地减低了误报率和漏报率。另外，文中利用滑动窗口计算各维特征的局部均值偏差，并剔除了异常数据，使其能够保证在实时动态变化的网络中的检测准确度。实验结果表明，该方法能有效地检测出多种异常，而且误报率相对较低。

参考文献

[1] Krugel C, Toth T, Kirda E. Service specific anomaly detection for network intrusion detection [C] //Proceedings of the 2002 ACM Symposium on Applied Computing. New York: ACM Press, 2002: 201 ~ 208.

[2] 郑黎明，邹鹏，韩伟红，等. 基于 Filter－ary－Sketch 数据结构的骨干网异常检测研究 [J]. 通信学报，2011，32(12)：151－160.

[3] Sengar H, Wang X, Wang H, et al. Online detection of network traffic anomalies using behavioral distance [C] //Proceedings of 2009 17th International Workshop on Quality of Service. New York: IEEE, 2009: 1－9.

[4] 诸葛建伟，王大为，陈昱，等. 基于 D－S 证据理论的网络异常检测方法 [J]. 软件学报，2006，17(3)：463－471.

[5] Portnoy L, Eskin E, Stolfo J. Intrusion detection with unlabeled data using clustering [C] //Proceedings of 2001 ACM CSS Workshop on Data Mining Applied to Security. Philadelphia: ACM Press, 2001: 5－8.

[6] Jiang Shengyi, Song Xiaoyu, Wang Hui, et al. A clustering－based method for unsupervised intrusion detections [J]. Pattern Recognition Letters, 2006, 27(7): 802－810.

[7] Kind A, Stoecklin M, Dimitropoulos X. Histogram－based traffic anomaly detection [J]. IEEE Transactions on Network and Service Management, 2009, 6(2): 110－121.

[8] Hareesh I, Prasanna S, Vijayalakshmi M, et al. Anomaly detection system based on analysis of packet header and payload histograms [C] //Proceedings of International Conference on Recent Trends in Information Technology. Piscataway: IEEE Computer Society, 2011: 412－416.

[9] Moore D, Voelker G M, Savage S. Inferring internet denial－of－service activity [C] //Proceedings of the 2001 USENIX Security Symposium. Washington, DC, 2001: 15.

[10] 龚俭，彭艳兵，杨望，等. 基于 Bloom Filter 的大规模异常 TCP 连接参数再现方法 [J]. 软件学报，2006，17(3)：434－444.

[11] Irfan U H, Sardar A, Hassan K, et al. NUST Traffic Dataset [EB/OL]. [2012－12－10]. http://wisnet.seecs.nust.edu.pk/projects/ENS/DataSets.html.

[12] Yan Ruoyu, Zheng Qinghua, Peng Weimin. Multi－scale entropy and renyi cross entropy based traffic anomaly detection [C] //Proceedings of 2008 11th IEEE Singapore International Conference on Communication Systems. Piscataway: IEEE Computer Society, 2008: 554－558.

(收稿日期：2013-03-18)