·安全技术· 文章编号:1000-3428(2018)01-0165-06 文献标志码:A 中图分类号:TP393

基于 ME-PGNMF 的异常流量检测方法

陈露露,郭文普,何 灏

(火箭军工程大学 信息工程系,西安 710025)

摘 要:由于部分网络异常对流量变化影响不明显,流量分析难以发现此类异常。传统基于主成分分析的网络异常流量检测方法追求全局最优解,对局部特征提取不充分,导致对连续异常不敏感,降低了异常流量的检测精度,且物理意义不明确。针对上述问题,在多维信息熵的基础上,提出梯度投影非负矩阵分解异常流量检测方法。将流量数据处理为多维特征熵矩阵,用梯度投影非负矩阵分解方法重构多维熵矩阵,分离出正常和异常子空间,采用多元统计过程控制方法中的 Q 图检测异常。实验结果表明,与以流量分析为基础的主成分分析方法、传统非负矩阵分解方法相比,该方法能更快、更准确地检测出连续异常,对流量变化不敏感的低速分布式拒绝服务攻击检测效果明显提高,对蠕虫攻击更加敏感。

关键词:网络流量;多维熵;异常检测;非负矩阵分解;子空间

中文引用格式:陈露露,郭文普,何 灏. 基于 ME-PGNMF 的异常流量检测方法[J]. 计算机工程,2018,44(1):165-170.

英文引用格式:CHEN Lulu, GUO Wenpu, HE Hao. Abnormal Traffic Detection Method Based on ME-PGNMF[J]. Computer Engineering, 2018, 44(1):165-170.

Abnormal Traffic Detection Method Based on ME-PGNMF

CHEN Lulu, GUO Wenpu, HE Hao

(Department of Information Engineering, Rocket Forces Engineering University, Xi'an 710025, China)

【Abstract】 Because some network anomalies have little effect on traffic flow, it is difficult to find such anomalies in traffic analysis. Traditional anomaly traffic detection method based on Principal Component Analysis (PCA) is not suitable for continuous local anomalies detection, and it can reduce the detection accuracy of abnormal flow and the physical meaning is not clear. Aiming at the above situation, an anomalous traffic detection method based on Multidimensional Entropy-Projected Gradient Non-negative Matrix Factorization (ME-PGNMF) is proposed. Firstly, the network traffic data is processed into multidimensional entropy matrix, then Projected Gradient Non-negative Matrix Factorization (PGNMF) is used to reconstruct the multi-dimensional entropy matrix, and the normal subspace and abnormal subspace are separated. Finally, the anomaly is detected by multivariate statistical process control chart Q. Experimental results show that the proposed method can detect the continuous anomaly faster and more accurately than the traditional Nonnegative Matrix Factorization (NMF) method based on the PCA method based on the flow analysis. The low-speed Distributed Denial of Serviee (DDOS) attack anomaly detection is not sensitive to the traffic change. Attacks are more sensitive.

【Key words】 network traffic; multidimensional entropy; abnormal detection; Non-negative Matrix Factorization (NMF); subspace

DOI:10.3969/j.issn.1000-3428.2018.01.028

0 概述

随着网络技术的发展和普及,网络流量的规模越来越大,类型也越来越多,网络安全问题日益突出。但不论何种网络异常都会对网络流量产生影响,因此,网络流量异常检测的研究越来越受到重视。网络流量异常检测是指通过对网络流量历史活动的观测建立网络流量的正常模式,和当前网络流量活动比较发现异常。其优点是可以发现新的异常,但也存在不少问题。

当前骨干网的网络活动特点是流量大、流速快、数据维数高,要进行数据检测,数据的降维处理成为一个必要的过程。文献[1]采用基于主成分分析(Principal Component Analysis, PCA)的子空间方法,利用 PCA 对数据降维的能力将流量矩阵分离为正常子空间和异常子空间,通过设置阈值,对大流量的

作者简介:陈露露(1993—),男,硕士研究生,主研方向为网络安全、通信工程;郭文普,副教授、博士;何 灏,硕士研究生。

收稿日期:2016-12-07 **修回日期**:2017-01-21 **E-mail**:879784162@qq.com

异常取得较好的检测效果。文献[2]同时考虑流量的时间和空间相关性,综合利用小波变换具有的多尺度建模能力和 PCA 具有的降维能力对正常流量进行建模,实现异常检测。但 PCA 方法仍存在较多问题:PCA 追求的是全局最优解,导致对局部特征提取不充分,影响检测精度;系数向量中存在负值,但流量不可能为负,可解释性差;PCA 方法的效果受主成分的选择和阈值设定的影响很大。鉴于以上问题,文献[3]将非负矩阵分解(Non-negative Matrix Factorization,NMF)方法用于异常流量检测。主要特点是采用非负子空间方法提取流量矩阵中的主要时变模式,构成正常子空间,得到重构矩阵和残余矩阵,然后应用 Q 图进行异常点检测。该方法有更强的特征提取能力,由于其非负性,可解释性也更强,实验表明比 PCA 方法在检测精度方面有明显提高。

传统的基于统计的网络流量异常检测方法多是对流量特性进行分析[4]。文献[5]总结了基于熵的异常流量检测的优点:在数据流上有高效的熵值计算算法,且有足够的精度保证;基于熵的检测方法提供了一种更细粒度的信息,能够检测出更加隐蔽的异常类型;采样对基于熵的检测方法精度影响并不明显;正常情况下熵值相对流量幅值更稳定。

文献[6]用增量投影非负矩阵分解对 NMF 方法进行改进,提高了 NMF 的效率。文献[7]虽利用了流量的熵值特征,但仅利用流量幅值的熵,流量数据利用不充分。为此,本文提出一种将特征熵和 NMF 相结合的方法对此类异常流量检测方法进行改进。

1 多维熵矩阵和 PGNMF 算法

1.1 多维熵矩阵

信息熵是信息论中用于度量信息量的一个概念,信息熵标志着所含信息量的多少,是对系统不确定性的描述。用信息熵来描述网络流量的特征,对流量数据预处理,不仅增强了对网络流量异常的检测能力,而且方便了流量异常的分类。

基于熵的异常检测系统的主要思想是:一旦有异常流量发生,总体流量的熵值应该会随之发生变化,通过熵值的变化能够检测出该异常。文献[8]把流量聚集成网络流,针对每个流在源/目 IP,源/目端口维度上分别计算熵值,然后采用 PCA 方法进行异常检测;文献[9]提出在高速 IP 网络上利用熵值来检测蠕虫爆发和流量异常,均取得较好的效果。文献[10]基于最大熵原理,通过比较当前分布和基准分布的差异来进行异常检测。文献[11]在多个不同维度上采用熵度量流量数据的分布特征,在每个时间窗口上把不同维度熵值序列排列成检测向量,提出多窗口关联检测算法,既降低了计算复杂度,也提高了检测效率。以上方法从不同层面都验证了熵特征在异常流量检测方面的良好性能。

把采集到的 Netflow 流量数据当作离散信息源,把数据中的各个属性看作是一组随机事件,就可以对它的信息熵进行分析。假设某一时段内目的 IP 的集合用 X 表示,i 表示有 i 个不同的目的 IP,n_i 表示第 i 个目的 IP 出现的次数:$X=\{n_i,i=1,2,\cdots,N\}$。那么该时段内目的 $IP(X)$ 的信息熵定义如式(1)所示,本文使用信息熵的指标有源/目的 IP、源/目的端口,如下:

$$H(X)=-\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\mathrm{lb}^{\left(\frac{n_i}{S}\right)},S=\sum_{i=1}^{N}n_i \tag{1}$$

不同指标熵的计算方法可参考式(1)。表 1 反映的是不同异常对流量各属性熵值的影响。

表 1 常见异常流量熵变化特征

异常类型	异常描述	异常熵特征
DOS	单源点对单(多)源点的攻击	源 IP 熵减小、目的 IP 熵增大
DDOS	多源点对单源点的攻击	源 IP 熵增大、目的 IP 熵减小
端口扫描	单源点向单源点短时间发送大量端口扫描信息	源 IP 熵减小、目的 IP 熵减小目的端口熵增大
蠕虫病毒	多个目的主机上的少部分目的端口被恶意探测	源 IP 熵减小、目的 IP 熵增大目的端口熵减小

为充分利用流量的时间和空间相关性,在多维特征熵的基础上构建多维熵矩阵。传统流量矩阵描述一个网络中所有源节点和目的节点(Origin-Destination, OD)对之间的流量情况,主要反映了目标网络各 OD 对之间流量分布情况。主要包括链路级、路由级和 PoP(Point of Presence)级流量矩阵。单一时刻的流量矩阵表示为一个 m 行的列向量:$(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$,其中。$m$ 表示 OD 对的数量。不同的列代表不同时刻。则在有 m 个 OD 对的网络中所有 n 个时刻的流量矩阵如式(2)所示,矩阵 $\boldsymbol{X}_{mn}$ 表示第 m 个 OD 对在 n 时刻的流量情况,多维熵矩阵就是将流量指标修改为源/目的 IP 熵,源/目的端口熵等指标得到四维熵矩阵。

$$\boldsymbol{X}_g=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&&\vdots\\x_{m1}&x_{m2}&\cdots&x_{mn}\end{pmatrix} \tag{2}$$

1.2 PGNMF 算法

NMF[12] 起源于主成分分析,1999 年 Lee 和 Seung 在 Nature 上发表了 NMF 算法,该算法是在矩阵中所有元素均为非负的条件下对其实现非负分解.具有可解释性、占用存储空间少等优点。NMF 可以定义为给定一个 $n\times m$ 阶非负矩阵 $\boldsymbol{V}_g$ 和一个正

整数 r,将矩阵 $\boldsymbol{V}_g$ 分解为 $n\times r$ 阶的矩阵 $\boldsymbol{W}_g$ 和 $r\times m$ 阶的矩阵 $\boldsymbol{H}_g$ 的乘积,如式(3)所示为得到近似的分解结果,需要定义一个目标函数:

$$\boldsymbol{V}_{m\times n}\approx \boldsymbol{W}_{m\times r}\times \boldsymbol{H}_{r\times n} \tag{3}$$

$$\min_{W,H} f(W,H)=\frac{1}{2}||\boldsymbol{V}-\boldsymbol{WH}||= \frac{1}{2}\sum_{i,j}\left(V_{ij}-\sum_{l=1}^{r}W_{il}H_{il}\right)^2 \tag{4}$$

来量化矩阵 $\boldsymbol{V}_g$ 与矩阵 $\boldsymbol{W}_g\times\boldsymbol{H}_g$ 的逼近程度,如式(4)所示。式(4)需满足矩阵 $\boldsymbol{W}_{ia}\geqslant 0,\boldsymbol{H}_{bj}\geqslant 0,\forall i,a,b,j$,是一个带约束的非线性规划问题。利用迭代的方法求解矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$。文献[3]给出了带约束条件下用拉格朗日算子解得的矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$ 的更新方程,如下:

$$\begin{cases} w_{ia}\leftarrow w_{ia}\dfrac{(\boldsymbol{VH}^{\mathrm{T}})_{ia}}{(\boldsymbol{WHH}^{\mathrm{T}})_{ia}} \\ h_{bj}\leftarrow h_{bj}\dfrac{(\boldsymbol{VW}^{\mathrm{T}})_{bj}}{(\boldsymbol{HWW}^{\mathrm{T}})_{bj}} \end{cases} \tag{5}$$

NMF 算法可描述如下:

1)初始化矩阵 $\boldsymbol{W}_{ia}\geqslant 0,\boldsymbol{H}_{bj}\geqslant 0,\forall i,a,b,j$。

2)设置基矩阵维数 r 和最大迭代次数 k,更新矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$,更新规则为式(5)。

3)循环执行式(5),至算法收敛。

但 NMF 算法复杂度高,不利于异常流量检测对实时性的要求,因此,采用投影梯度(Projected Gradient, PG)优化方法降低 NMF 迭代的复杂度。PGNMF 方法具物理意义明确、稀疏性好、以及耗时少等特点,但梯度投影法使用另外的迭代规则。文献[9]提出的一种梯度投影法,使迭代的复杂度大为降低且更新得到的解收敛。算法基本思想如下:

将式(4)改写为如下形式:

$$\begin{cases} \min\limits_{W,H}\bar{f}(\boldsymbol{H})=\dfrac{1}{2}||\boldsymbol{V}-\boldsymbol{WH}||=\dfrac{1}{2}\sum\limits_{i,j}(V_{ij}-(\boldsymbol{WH})_{ij})^2 \\ H_{bj}\geqslant 0,\forall b,j \end{cases} \tag{6}$$

$$\begin{cases} \min\limits_{W,H}\bar{f}(\boldsymbol{W})=\dfrac{1}{2}||\boldsymbol{V}-\boldsymbol{WH}||=\dfrac{1}{2}\sum\limits_{i,j}(V_{ij}-(\boldsymbol{WH})_{ij})^2 \\ W_{ib}\geqslant 0,\forall i,b \end{cases} \tag{7}$$

其中,$\bar{f}(\boldsymbol{H})$、$\bar{f}(\boldsymbol{W})$分别为固定矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$ 的函数。

以式(6)为例,梯度投影方法是对式(6)进行边界优化。假设 k 是迭代的索引,$\overline{H}$ 是得到的当前最优解,则 PG 方法通过 $\tilde{H}\equiv P[\overline{H}-\alpha_k\nabla\overline{f}(\overline{H})]$ 更新 $\overline{H}$ 为 $\tilde{H}$,其中梯度投影函数 $P[x]=\max(0,x)$,取 0 ~ x之间的最大值,α_k 为梯度投影的步长。步长选择是 PG 方法中最耗时的部分,具体更新算法可参见文献[13]。

更新 $\tilde{H}\equiv P[\overline{H}-\alpha_k\nabla\overline{f}(\overline{H})]$所需的时间复杂度为 $O(tnmr)$,若存在 t 个 $\overline{H}$,则时间复杂度增为 $O(tnmr)$,因此 Lin 将步长 α_k 的迭代条件改为:

$$(1-\sigma)\langle\overline{f}(\overline{H}),\tilde{H}-\overline{H}\rangle+\frac{1}{2}\langle\tilde{H}-\overline{H},(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})(\tilde{H}-\overline{H})\rangle\leqslant 0 \tag{8}$$

此时,计算复杂度主要集中在$(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})(\tilde{H}-\overline{H})$,为 $O(nr^2)$,复杂度因此大为降低。综上,PGNMF 算法步骤如下:

1)初始化矩阵 $\boldsymbol{W}_{ia}\geqslant 0,\boldsymbol{H}_{bj}\geqslant 0,\forall i,a,b,j$。

2)设置基矩阵维数 r 和最大迭代次数 k,通过迭代条件式(8),利用梯度投影方法求解式(6)、式(7),得到矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$ 的更新关系。

3)重复步骤(2),至算法收敛,得到矩阵 $\boldsymbol{W}_g$ 和 $\boldsymbol{H}_g$。

2　基于 ME-PGNMF 异常流量检测模型

基于 ME-PGNMF 的异常流量检测方法主要分为构建多维熵矩阵、获取残差矩阵、异常点检测。如图 1 所示为异常流量检测的基本流程。

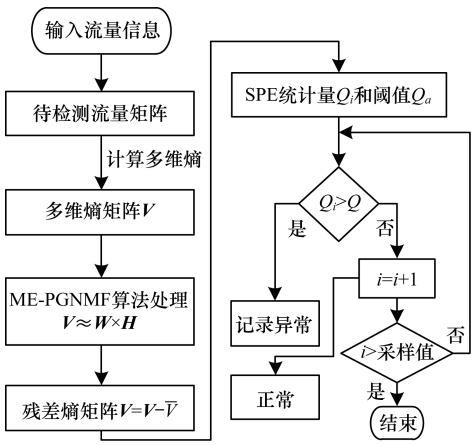

图 1　异常流量检测基本流程

2.1　多维熵矩阵构建

若矩阵 $\boldsymbol{X}_{m\times n}$ 表示待检测的目标网络流量矩阵,则 m 表示目标网络的 OD 对的数量,n 表示检测时间段内的采样周期数。根据信息熵定义,求出网络流量在 n 个周期内源/目的 IP、源/目的端口 4 个属性的熵值序列,并标准化处理,得到 4 个大小为 $m\times n$ 的标准化熵矩阵。将 4 个维度矩阵依次按行相接得到一个 $4m\times n$ 的多维熵矩阵,用矩阵 $\boldsymbol{V}_{4m\times n}$ 表示。矩阵 $\boldsymbol{V}_{4m\times n}$ 的列向量表示某一检测周期内不同 OD 对 4 种属性熵值的变化,行向量表示某一 OD 对不同周期某一属性熵值变化。

2.2 残差矩阵获取

若将多维熵矩阵中第 i 个位置的向量看做是 d 维空间的点,每一个点的状态特征用多维熵特征表示。多维熵矩阵可以用 r 维的子空间表示,满足条件的 r 维子空间为多维熵矩阵的特征子空间,而 r 维子空间可以通过 PGNMF 算法进行构建。选取子空间维数 r 和迭代次数 k,通过构建 r 维特征子空间,得到残差距阵。残差矩阵的获取分 2 个步骤:提取正常成分和获取残余成分。

首先提取正常成分,对多维熵矩阵 $\boldsymbol{V}$ 进行 PGNMF 算法分解之后可以获得矩阵 $\boldsymbol{W}$ 和矩阵 $\boldsymbol{H}$,其中基矩阵 $\boldsymbol{W}_g=(w_1,w_2,\cdots,w_r)$ 的每一列代表的是 r 维子空间的基向量,系数矩阵 $\boldsymbol{H}_g=(h_1,h_2,\cdots,h_n)$ 中的每一列表示的是 r 维子空间的系数向量,而每一个基向量捕获的是目标网络各 OD 对的一种状态模式。矩阵 $\boldsymbol{V}_g\approx\boldsymbol{W}_g\times\boldsymbol{H}_g$ 代表的是目标网络的正常状态成分。

得到正常流量状态后,通过当前流量状态与正常状态之差,即矩阵 $\tilde{\boldsymbol{V}}_g=\boldsymbol{V}_g-\overline{\boldsymbol{V}}_g$,得到残余成分。

2.3 异常点检测

得到残差矩阵后,如果某周期内发生了异常,其对应的残差向量也会发生明显变化。本文引入 Q 统计量来监测残差向量的变化,Q 统计量的时序图亦称平方预测误差(Squared Prediction Error, SPE)图[14],是多元统计过程控制图的一种,主要监测输入多变量的数据结构是否变化,残差向量中包含源/目的 IP、源/目的端口熵 4 个变量,每个残差向量代表一个采样点。

Q 统计量在 i 时刻的值 Q_i 是个标量,用残余向量中所有元素的平方和(SSE)来衡量,当检验水平为 α 时,控制限 Q_α 可按式(9)计算:

$$\begin{cases} Q_\alpha=\theta_1\left[\dfrac{c_\alpha h_0(2\theta_2)^{1/2}}{\theta_1}+1+\dfrac{\theta_2 h_0(h_0-1)}{\theta_2}\right]^{1/h_0} \\ \theta_i=\sum\limits_{j=k+1}^{n}\lambda_j^i,i=(1,2,3) \\ h_0=1-2\theta_1\theta_3/(3\theta_2^2) \end{cases} \tag{9}$$

其中,λ_j 为多维熵矩阵 $\boldsymbol{V}$ 第 j 个特征值,c_α 是正态分布中 $1-\alpha$ 分位数,α 通常取 0.001。若 $Q_i<Q_\alpha$ 说明该时刻流量正常,否则为异常流量。

分析以上 3 个步骤可以发现基于 ME-PGNMF 的异常流量检测方法有以下 3 个关键问题:1)PGNMF算法的效果受初始参数 r 和 k 影响较大,如何选取合理的参数保证检测效果。2)r 维特征子空间是否能还原出流量的正常状态模式,从而保证得到的残余矩阵包含了异常状态模式。3)用 Q 统计量对残余向量进行多元统计过程控制是否能检测出异常。对于问题 1)可通过矩阵还原性来选取参数,问题 2)一方面取决于矩阵还原性,同时取决于检测效果,而问题 3 主要取决于检测效果。综上,在进行实验设计和分析时要从以上 3 个问题进行综合考虑,不仅要分析最终的检测效果,还要关注过程中的矩阵还原性指标。

3 实测数据实验及结果分析

3.1 实验数据与方法

为了验证分析 ME-PGNMF 算法在异常流量检测方面的有效性,实验实测数据来自 Abilene 骨干网在 2009 年 3 月 1 日—2009 年 3 月 7 日一周所采集的流量矩阵数据。该网络是美国的 IP 骨干网络,每个节点对应一个城市,由 12 个节点组成,共有 $12\times12=144$ 条 OD 流,所有节点通过配置 NetFlow 来收集真实 OD 流。其中,采集间隔为 5 min,这样一周所采集的次数为 $7\times24\times60/5=2\,016$,而 OD 流的数目为 144 个,因此,该 OD 流矩阵为一个 144×2016 的矩阵。在 OD 流矩阵的基础上,计算源/目的 IP、源/目的端口 4 个同样大小的熵矩阵,标准化处理后,按行相接得到 $576\times2\,016$ 的多维熵矩阵。表 2 所示为本文实验的具体数据集形式。

表 2 不同算法的流量数据形式

数据源	算法	测度	矩阵描述	数据形式
Abilene 数据集	PCA	数据流	$144\times2\,016$	流量大小
(2009.3.1—	NMF	数据流	$144\times2\,016$	流量大小
2009.3.7)	PGNMF	数据流	$144\times2\,016$	流量熵
	ME-PGNMF	数据流	$576\times2\,016$	多维熵

但 Abilene 数据集中并无明显异常,因此采用人为注入异常流量的方法进行实验。为方便和 PCA、NMF 等方法对比,采用和文献[3]相似的异常注入方式。在实验中,采样点之间间隔为 5 min,异常注入过程如下:从第 300 个采样点到第 800 个采样点期间,每隔 50 个采样点轮流依次注入一次低速 DDOS 攻击和一次 DDOS 攻击,并且 2 种攻击都持续 30 min(持续 6 个采样点);从第 1 000 个 ~ 第 1 500 个采样点里,每隔 50 个采样点注入一次端口扫描,攻击持续时间为 30 min(持续 6 个采样点);从第 1 700 个 ~ 第 1 900 个采样点期间,持续注入蠕虫攻击,包括个感染到爆发的完整过程(持续 200 个采样点)。

3.2 参数分析与选择

非负矩阵分解算法的一个非常重要的参数是基矩阵维数 r 的选择,参数 r 主要影响分解后的矩阵对原矩阵的还原性,是影响最终检测结果的重要因素。且参数 r 的选择受数据集特性影响较大。迭代次数 k 虽然在足够大的情况下能确保算法收敛,但 k 值过大会使算法的复杂性增大。鉴于以上情况以及数据

集的相似性,对基于流量矩阵的 NMF 方法,参数选择参考文献[3],对基于流量熵矩阵的 PGNMF 方法,参数选择参考文献[15],本文重点分析基于多维熵矩阵的 PGNMF 方法的参数选择。

矩阵的还原性通常用式(10)来衡量:

$$d=\frac{1}{2}\sum_{ij}[V_{ij}-(\boldsymbol{WH})_{ij}]^2 \tag{10}$$

为确定基矩阵维数 r 和合理的迭代次数 k,为确保 d 收敛,首先设置最大的迭代次数 k 为 500,r 从 2 递增至 20,寻求最优矩阵还原性的 r。。实验对象为标准化后的 576 × 2 016 的多维熵矩阵,实验结果如图 2 所示。

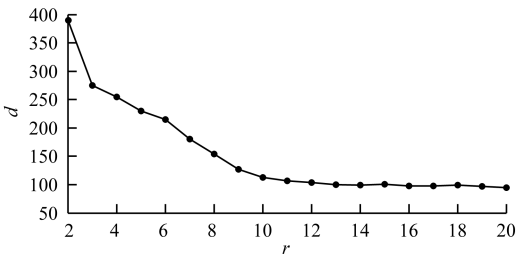

图 2　矩阵还原性 d 与维数 r 的关系

从图 2 和图 3 可以看出,基矩阵维数 r 设为 12 时,矩阵还原性 d 已基本稳定即 d 收敛,且此时收敛所需时间 t 还没有明显增加。结合图 4,d 收敛时所需的迭代次数 k 均未超过 200,因此,最大迭代次数 500 未对实验造成影响。综合以上结果选取参数 r 为 12,k 为 200。

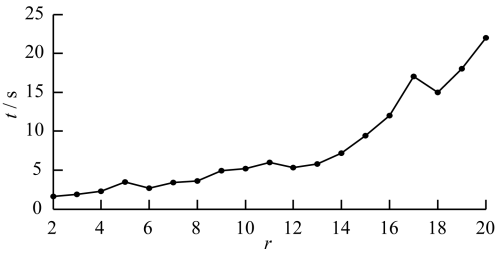

图 3　d 收敛时,收敛时间 t 与维数 r 的关系

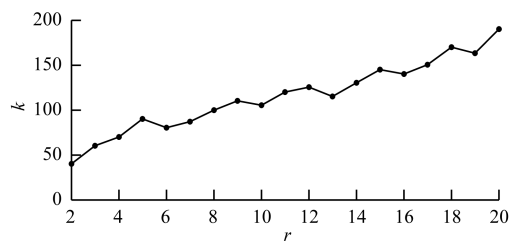

图 4　d 收敛时,不同维数 r 与迭代次数 k 的关系

3.3　实验结果及分析

对基于 NMF 的检测算法选择合适的基矩阵维数 r 和迭代次数 k 后,按表 2 算法与实验数据形式得到图 3 和表 3、表 4 的实验结果。图 3 中直线代表控制限 Q_α,α 取 0.001,圆圈代表残余向量的 Q_i。表 3 检测率指标参考文献[16],为检测到的异常点与注入异常点的比,表 4 中误报率为注入异常时段内正常采样点报为异常的次数与正常采样点的比。

表 3　不同方法对注入异常的检测率　%

方法	低速 DDOS	DDOS	端口扫描	蠕虫攻击
PCA	2.3	78.4	81.0	56.2
NMF	13.5	86.9	82.3	70.6
PGNMF	16.4	91.2	89.5	74.0
ME-PGNMF	93.4	95.7	96.8	91.4

表 4　不同方法对注入异常的误报率　%

方法	低速 DDOS	DDOS	端口扫描
PCA	0.0	12.8	17.4
NMF	3.6	9.7	7.9
PGNMF	2.7	7.5	5.8
ME-PGNMF	4.7	9.6	6.7

通过分析表 3 和图 5 ~ 图 8 可以发现对于低速 DDOS 攻击,PCA 方法检测效果最差,检测率接近于 0,NMF 和 PGNMF 方法检测效果也不明显检测率不足 20% ,分析原因主要是此类攻击并未明显引起流量变化,而这 3 种方法都是以流量为基础。对于端口扫描和普通 DDOS 攻击,几种方法都能不同程度的检测出来,其中,以 ME-PGNMF 方法效果最好,原因是该方法既有 NMF 方法局部特征提取能力强的优势,又以多维熵为基础,对异常流量的分辨能力有很大提高。最后,比较对蠕虫攻击的检测可以发现,PCA 方法对攻击最不敏感,在第 1 700 个采样点攻击已出现,图 5 直到第 1 780 个采样点左右才有检测到的迹象,图 6 和图 7 中虽然在第 1 700 采样点已能检测到异常但并不稳定,依然有很多漏报点发生,直至第 1 750 采样点才稳定报出异常,只有 ME-PGNMF 方法最快对蠕虫攻击做出反应,且漏报点很少,分析原因,与蠕虫攻击在起始阶段对流量变化影响不大有关。

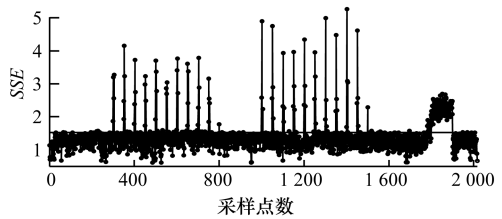

图 5　PCA 方法异常流量检测效果

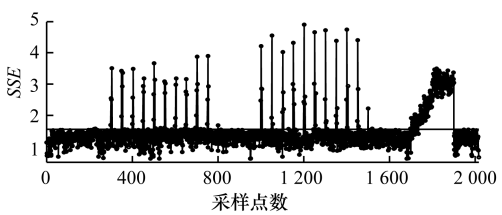

图 6　NMF 方法异常流量检测效果

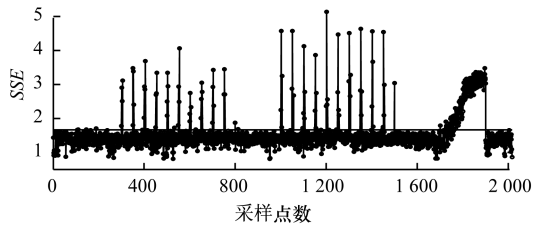

图 7 PGNMF 方法异常流量检测效果

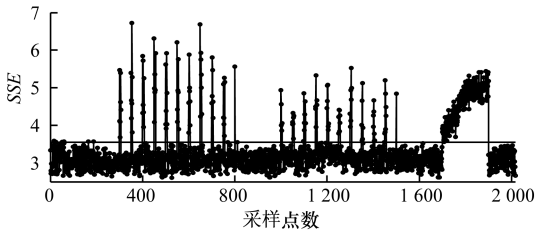

图 8 ME-PGNMF 方法异常流量检测效果

为准确评价几种方法对注入异常的检测效果，分析表 4 的误报率可以看出，基于流量的方法对低速 DDOS 攻击不敏感，检测率和误报率都很低，对其他攻击流量，ME-PGNMF 方法检测率有明显提高，误报率也没有明显上升。综合考虑 3 种攻击，ME-PGMF 方法检测率有明显提高，同时误报率并无明显提升。对于蠕虫攻击，由于每一个采样点都注入异常，因此不存在把正常采样点报为异常点，即误报问题。

3.4 参数选择对实验结果的影响

在 3.2 节中分析了 ME-PGNMF 方法中参数选择对算法的影响，为直观体现参数 r 和 k 对实验结果的影响，通过设计不同的参数得到图 9 ~ 图 10 所示的实验结果。可以看出，实验结果与参数分析基本一致。r 为 12，k 为 120（小于实验预设值 200）都达到了实验的最佳效果，如果参数变大虽能达到检测效果，但运算的时间复杂度就会增加。

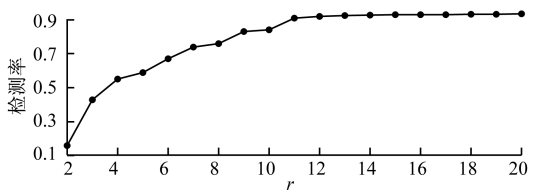

图 9 k 为 200 时，r 与检测率的关系

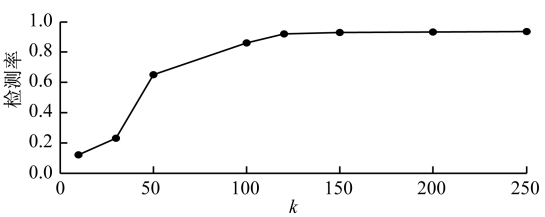

图 10 r 为 12 时，k 与检测率的关系

4 结束语

利用熵特征对异常流量的敏感性和 PGNMF 算法局部特征提取能力强、收敛速度快的特性，本文提出基于多维熵和 PGNMF 的网络异常流量检测方法。实验结果表明，与基于流量的检测方法相比，该方法对低流量异常更敏感，有利于发现隐蔽的流低速攻击，可以在蠕虫爆发的早期阶段及时发现，提高了检测率。下一步将改进数据输入形式，通过增量或其他方法，达到在线检测的目的。

参考文献

[1] LAKHINA A, CROVELLA M, DIOT C. Mining Anomalies Using Traffic Feature Distributions[J]. Conference on Applications, 2005, 35(4): 217-228.

[2] 钱叶魁，陈 鸣，叶立新，等. 基于多尺度主成份的全网络异常检测方法[J]. 软件学报，2012，23(2)：361-377.

[3] 魏祥麟，陈 鸣，张国敏. NMF-NAD：基于 NMF 的全网络流量异常检测方法[J]. 通信学报，2012，33(4)：54-61.

[4] 王 浩. 针对 TCP 的低速 DDoS 解析及防御策略[J]. 计算机工程，2009，35(13)：122-124.

[5] 郑黎明. 大规模通信网络流量异常检测与优化[D]. 长沙：国防科学技术大学，2012.

[6] 柏 骏，夏靖波，吴吉祥，等. ODA-IPNMF：一种在线全网络流量异常检测方法[J]. 哈尔滨工业大学学报，2015，47(5)：104-109.

[7] 王晓鸽. 基于 PGM-NMF 的网络流量异常检测研究[J]. 电子科技，2014，27(5)：175-178.

[8] ANUKOOLL, MARK C, CHRISTOPHE D. Mining Anomalies Using Traffic Feature Distributions[C]// Proceedings of ACM SIUCOMM'05. New York, USA: ACM Press, 2005: 217-228.

[9] WAGNER A, PLANNER 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 D. C., USA: IEEE Press, 2005: 172-177.

[10] YU Gu, ANDREW M, DON T. Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation[C]// Proceedings of the 5th ACM SIUCOMM Conference on Internet Measurement. New York, USA: ACM Press, 2005: 345-350.

[11] 郑黎明，邹 鹏，韩伟红，等. 基于多维熵值分类的骨干网流量异常检测研究[J]. 计算机研究与发展，2012，49(9)：1972-1981.

[12] LEE D, SEUNG H S. Learning the Parts of Objects by Non-negative Matrix Factorization[J]. Nature, 1999, 401(6755): 788-791.

[13] LINC J. Projected Gradient Methods for Non-negative Matrix Factorization[J]. Neural Computation, 2007, 19(10): 2756-2779.

[14] 王兆军，邹长亮，李忠华，等. 统计质量控制图理论与方法[M]. 北京：科学出版社，2013.

[15] 王晓鸽. 基于流量矩阵的网络入侵检测研究[D]. 兰州：兰州交通大学，2014.

[16] 许 倩，程东年. 基于层次聚类的网络流量异常分类算法[J]. 计算机工程，2012，38(23)：131-136.

编辑 刘 冰