基于多尺度低秩模型的网络异常流量检测方法

程国振，程东年，俞定玖
（国家数字交换系统工程技术研究中心，河南 郑州 450002）

摘 要：现有刻画流量异常检测所需的流特征集通常是高维的，增加了检测和分类的复杂度。通过研究发现网络中异常通常是稀疏性分布的，单个异常仅仅表现在低维流特征中。基于这一现象提出了一种异常流量检测模型—多尺度低秩 (MRLR, multi-resolution low rank)模型，该模型能够动态筛选出“合适的”特征集并准确分类异常。基于人工标记的实际网络流量异常和注入异常的数据集验证结果表明：MRLR 对特征集的缩减率可达 10%以下；并且基于 MRLR 的分类算法复杂度为 $O(n)$。

关键词：异常检测；特征选择；多尺度分析；低秩分布
中图分类号：TP393.08 **文献标识码**：A **文章编号**：1000-436X(2012)01-0182-09

Network traffic detection based on multi-resolution low rank model

CHENG Guo-zhen, CHENG Dong-nian, YU Ding-jiu
(National Digital Switching System Engineering &Techno gical R&D Center, Zhengzhou 450002, China)

Abstract: Because network traffic was usually characterized by its higher-dimensional features, related detectors and classifiers for identifying traffic anomalies were suffering the increased complexity. Several key observations given by existing studies showed that network anomalies were distributed typically in a sparse way, and each of anomalies was essentially characterized by its lower-dimensional features. Based on this important finding, a novel model detecting traffic anomalies—multi-resolution low rank (MRLR) was developed. The proposed MRLR allowed us to dynamically filter the “proper” feature sets and then to classify anomalies accurately. The validation shows that MRLR can accurately reduce the dimensions of flow features to lower than 10%, on the other hand, the complexity of MRLR-classifiers is $O(n)$.

Key words: anomaly detection; feature filtering; multi-resolution analysis; low-rank distribution

1 引言

在大型 ISP (internet service provider)网络或者企业网中检测异常是困难的。首先，存在各种各样的异常。异常可能来自具有恶意企图的活动(如扫描，分布式拒绝服务攻击(DDoS, distributed denial of service))，也可能是误操作和网络元素的故障（例如，链路故障，路由问题，测量设备的缓冲溢出等），甚至来自诸如不寻常的大块文件传输(如 FTP(file transfer protocol))和突发访问量。其次，存在高维流量特征表征异常。在检测过程中，如果所选特征是低维的，则不足以描述流量及其含有异常的特性；如果所选特征是高维的，则增加了检测和分类模块的计算复杂度。因此如何根据实际流量动态选择合

收稿日期：2010-12-20；修回日期：2011-09-30
基金项目：国家高技术研究发展计划（“863”计划）基金资助项目（2009AA01A346, 2009AA01A334, 2008AA01A325, 2008AA01A326）；国家“十一五”科技支撑计划基金资助项目（2008BAH37B02）
Foundation Items: The National High Technology Research and Development Program of China (863 Program)(2009AA01A346, 2009AA01A334, 2008AA01A325, 2008AA01A326); The Science-Technology Support Project of the National “Eleventh Five-Year-Plan” of China (2008BAH37B02)

适的流特征检测异常是研究人员面临的挑战。

早期的检测技术大多数以流量的大小作为指标，例如统计的分组数或字节数[1,4]。Nychis 等[3]分析了流大小和度数分布的信息熵，可以检测到小流量的异常，而采用 IP 地址和端口号的熵却不能。文献[4]将 D-S 证据理论应用到异常检测中用以融合不同流特征数据，通过得到的基本信任分配函数检测异常，但也存在输入特征维数灾难(curse of dimensionality)的风险。Lakhina[5]通过输入流分布(例如 IP 地址和端口号)的剩余熵分类由 PCA[6] (principal component analysis)检测器识别的异常。当流量中含有大异常时，PCA 会偏离正常模式，而剩余熵又是 PCA 检测器的内部变量，其准确性受到 PCA 的固有缺陷的限制[7]。徐琴珍等[8]提出了一种分层的分层支持向量机(HSVM, hierachical support vector machine)的检测机制用以建立稳定的学习模型，并以更精简的形式表示特征。

Leland 等[9]和 Paxson 等[10]分析不同网络的流量均发现流量具有自相似特性。从物理意义上看，自相似过程的持续表现为长相关性(LRD, long range dependence)，也称之为多尺度行为特性。Paul Barford 等[1]首次将小波分析引入异常检测不同尺度上的异常。Garcia R C 等[11]利用小波分析检测短时异常。任勋益等[12]利用流量的自相似性检测 DDoS 攻击。文献[13]基于流量的宏观多分性特性检测多种异常。这些方法为异常检测的研究提供了坚实的理论和时间基础，同时也遇到了以下问题：第一，仅能检测引起流量结构变化的异常，对于小异常无能为力；第二，均采用小波分析进行流量的多分辨率分析难以达到实时性的要求。

基于上述分析，本文提出了一种多尺度低秩(MRLR, multi-resolution low rank)模型。基于 MRLR 模型的递归缩减特征 (RRF, recursive reducing features)算法通过动态地学习检测异常的“合适”的流特征集降低流特征的维数。最后将学习到的低维特征集输入聚焦分类算法(FCA, focused classification algorithm)中进行异常分类。实验表明，RRF 对高维流特征集的缩减率达到 10%下；基于 FCA 的简单分类器满足了实时性的要求。当异常密度较稀疏时，该方法能够准确地检测到异常，随着异常密度增加但不超过一定范围时(第 4 节实验显示不超过 30%)，其性能仍可保持在较高的检测率。

本文按下面的结构进行组织：第 2 部分给出了多尺度低秩模型的定义及论证；第 3 部分给出了特征选择算法及异常分类算法的思想；第 4 部分给出了基于 MRLR 模型的应用的仿真实验；第 5 部分是结束语。

2　多尺度低秩模型

本节根据异常流量的分布及表现行为提出了多尺度低秩(MRLR)模型。该模型成立的前提条件是假设异常在网络中的分布具有低秩特性。经研究发现以下 2 个事实。

1) 异常在网络链路中的分布是稀疏的。某一时刻，网络中的绝大多数链路处于正常情况，异常有局部集聚特性，仅影响少数链路。4.2 节中的图 5 验证了这一事实。

2) 单个异常仅仅反映在低维流特征上[14]。如表 1 所示，常见异常的发生仅导致低维流特征表现为异常。例如监视 H(dest_IP)，flow counts 等特征即可识别 DDoS 流量。

表 1　常见的异常及其所影响的流特征

异常	描述	用于检测的流特征
Alpha	不寻常的点对点高速数据传输，涉及到网络中的某源目的 IP	# of Packages,# of Bytes
DDoS	针对某个目的地址的分布式拒绝服务攻击	H(dest_IP), flow counts, 不重复# of src_IP，# of packages
Network scan	针对某一端口扫描整个网络	H(src_ip), H(dest_port), flow counts, # of packages
Port scan	扫描网络中易攻击的端口	H(src_ip)，H(src_port)，不重复# of src_port ，# of packages
Flash crowd	不正常的大量资源或服务的请求	Flow counts, H(dest_ip), 有时 H(dest_port)

其中，H(?)表示信息熵。

2.1　模型定义

假设可标识异常的流特征集为 $F=\{f_1(t), f_2(t),\cdots, f_n(t)\}$，$f_i(t)$ $(i=1,2,\cdots,n)$是在相同时间间隔内统计得到的时间序列，表征特征 i 的时间序列。考虑到流量的尺度特性，在建立低秩模型前对流特征进行 EMD 多尺度分解，凸显并稀疏化异常。对特征序列 $f_i(t)$进行 p 层经验模态分解去除余量等趋势分量得到 $f_{ij}(t)$ $(i=1,2,\cdots,n; j=1,2,\cdots,p)$。将时间 t 离散化，并取检测时间窗口长度为 M，则特征 f_{ijk} $(i=1,2,\cdots,n; j=1,2,\cdots,p; k=1,2,\cdots,M)$表示在时间段 k，特征 i 的第 j 个尺度分量的值。如果将每个尺度分量作为流特征，则可重写为 f_{ij} $(i=1,2,\cdots, N; j=1,2,\cdots,M)$，其

中，$N=np$。

定义1 （掩码矩阵(mask matrix)）设流特征矩阵 $\boldsymbol{F}(i,j)$ (i=1,2,…,N；j=1,2,…,M)，定义掩码矩阵 $\boldsymbol{Q}$ 为

$$\boldsymbol{Q}=\begin{cases}0, & F(i,j)\text{ 无异常}\\ 1, & \text{其他}\end{cases} \tag{1}$$

如果掩码矩阵 $\boldsymbol{Q}$ 是稀疏化的，则称 $\boldsymbol{F}(i,j)$中的异常具有稀疏分布特性。筛选后的流特征矩阵为 $\boldsymbol{F}'=\boldsymbol{Q}.*\boldsymbol{F}$。其中，.*表示相同大小的矩阵对应元素相乘，例如 $\boldsymbol{A}=\boldsymbol{C}.*\boldsymbol{B}$ 表示 $\boldsymbol{A}(i,j)=\boldsymbol{B}(i,j)\boldsymbol{C}(i,j)$。

这里的稀疏性是指某一向量只有很少非零元素，或者向量仅有少数元素具有相对较大值，其余元素的值很小。从矩阵的角度分析，低秩与稀疏性相似，因为低秩矩阵的奇异谱是稀疏的[15]。值得说明的是，传统检测异常的方法均假设异常改变了正常流量的结构特性。MRLR 模型并不与之矛盾，相反，正是传统检测方法的补充。MRLR 模型适用于异常稀疏的情况，以至于背景流量淹没了异常流量。

2.2 模型论证

设表征各类异常的流特征全集是 n 维集合 F，某时刻出现一个异常，该异常所影响的特征子集是 l 维集合 F'。基于事实 1)、2)本节给出以下结论。

定理1 假设 $K(?)$为任一计算特征偏离正常模式的检测器函数，矩阵 $\boldsymbol{A}=K(F)$，$\exists e>0$，使

① 当 $\boldsymbol{A}(i,j)>e$ 时，$\boldsymbol{A}(i,j)=1$

② 当 $\boldsymbol{A}(i,j)<e$ 时，$\boldsymbol{A}(i,j)=0$

则 $\boldsymbol{A}$ 是可低秩估计矩阵。

推论1 流特征矩阵经过 EMD 多尺度分解增强了定理1中的稀疏性。

推论2 假设网络中没有大规模爆发任何不可控异常的情况下，掩码矩阵 $\boldsymbol{Q}$ 是稀疏的，至少是行稀疏的。

证明 低秩矩阵估计(low rank approximation)的一个基本工具是奇异值分解(SVD, singular value decomposition)。矩阵 $\boldsymbol{A}$ 可以被分解为 $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$。其中，$\boldsymbol{V}^{\mathrm{T}}$ 是矩阵 $\boldsymbol{V}$ 的转置，$\boldsymbol{U}$ 是 $N\times N$ 的归一化矩阵(也就是 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}=\boldsymbol{I}$)，$\boldsymbol{V}$ 是 $M\times M$ 的归一化矩阵(即 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{I}$)，$\boldsymbol{S}$ 是一个 $N\times M$ 包含 $\boldsymbol{A}$ 奇异值 s_i 的对角矩阵。一般奇异值降序排列即 $s_i\geq s_{i+1}$。矩阵的秩是线性独立的行或者列，大小等于非零奇异值的个数。

为了进一步理解 SVD 在矩阵估计的应用，考虑以下 SVD 的解释。矩阵 $\boldsymbol{S}$ 是对角矩阵，所以 $\boldsymbol{A}$ 的 SVD 可以表示为

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}=\sum_{i=1}^{\min(N,M)}s_i u_i v_i^{\mathrm{T}}=\sum_{i=1}^{\min(N,M)}s_i\boldsymbol{A}_i \tag{2}$$

其中，u_i 和 v_i 分别是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的第 i 列，矩阵 $\boldsymbol{A}_i$ 的秩1矩阵。$\boldsymbol{A}$ 可以通过在 SVD 中只保留 r 个最大的奇异值产生秩 r 的估计矩阵 $\hat{\boldsymbol{A}}$，即

$$\hat{\boldsymbol{A}}=\sum_{i=1}^{r}s_i\boldsymbol{A}_i \tag{3}$$

其中，$\hat{\boldsymbol{A}}$ 是最佳秩 r 矩阵估计，估计误差可用 F 范数表示，所谓 F 范数是指对于任一矩阵 $\boldsymbol{Z}$，其 F 范数为 $\|\boldsymbol{Z}\|_{\mathrm{F}}\sim\sqrt{\sum_{i,j}\boldsymbol{Z}(i,j)^2}$。SVD 分解的截断是以下优化问题的解：

$$\begin{aligned}&\text{minimize}\ \ \left\|\boldsymbol{A}-\hat{\boldsymbol{A}}\right\|_{\mathrm{F}}\\ &\text{subject to}\ \ \mathrm{rank}(\hat{\boldsymbol{A}})\leq r\end{aligned} \tag{4}$$

由于 $\boldsymbol{A}$ 是稀疏矩阵，因此 $r\ll\min(N,M)$，$\hat{\boldsymbol{A}}$ 是 $\boldsymbol{A}$ 的低秩估计矩阵。

表1给出了常见的异常及其影响的流特征。每类异常的特征子集 F'仅有少数流特征可有效检测相应的异常 $l\leq 5$。另外，从统计的角度分析，设 v_i 代表异常发生时流特征 f_i 的归一化的偏移的绝对值。那么 v_i 在所有流特征变化所占的比率定义为

$$P(f_i)=v_i\left(\sum_{i=1}^{n}v_i\right)^{-1} \tag{5}$$

图1基于 KDD CUP1999 数据集给出了 $P(f_i)$ 的累积分布，横轴表示流特征的编号，纵轴表示流特征的偏移。可以看出：各个异常曲线仅在离散的几个点处出现大的跳跃，这证明跳跃点对应的流特征的偏移在所有的流特征中占了主导地位。此现象从另一方面验证了异常对流特征的影响的稀疏性。

当对特征集进行 p 层多尺度分解后，该异常被映射到某一或者某 q 个尺度分量上(其中，$p\geq q$)，某异常所影响的特征子集仍是 lq 维集合 F'，而特征全集 F 成为 np 维。由于 $n\gg l$，可得 $np\gg lq$。分解前的稀疏度 $S_a=l/n$，分解后的稀疏度 $S_b=lq/np$，由于 $p\geq q$，那么 $S_a\geq S_b$，因此推论1成立。

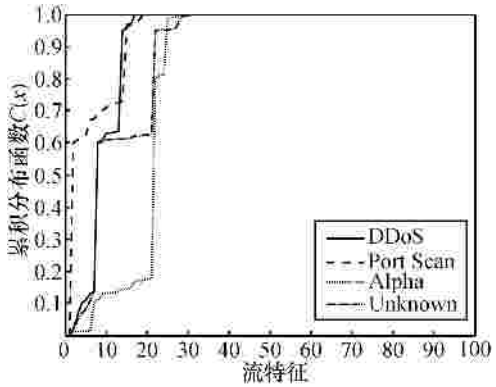


图 1　流特征的偏移累积分布

对于推论 2，由于网络中没有爆发大规模不可控异常，因此网络中的异常也是稀疏的；由定理 1 和推论 1 可知，因为表征单个异常的特征子集是低维的，即 $np \gg lq$，所以矩阵 $\boldsymbol{Q}$ 是行稀疏的。设 D_i 表示异常引起第 $i(i=1,2,\cdots,l)$个流特征变化的持续时间的长度。相对于观察的时间窗口的长度 M，如果对于$\forall i, M \gg D_i$，矩阵 $\boldsymbol{Q}$ 也是列稀疏的，故 $\boldsymbol{Q}$ 是稀疏的，否则不是列稀疏的。

3　基于 MRLR 模型的异常检测

本节基于 MRLR 模型的相关结论提出了一种异常特征筛选算法，实验证明该算法能快速、准确地检测异常特征。

3.1　动态特征选择算法

为了提高检测器的检测效率，降低分类器的复杂性，基于 MRLR 模型给出了一种动态学习“最佳”特征的算法——递归缩减特征 (RRF, recursive reducing features) 算法。RRF 算法基于定理 1，思想是将矩阵 $\boldsymbol{F}$ 稀疏最大化，即最小化异常特征集 A_f。一个向量的最稀疏问题可以抽象为 l^0 范数最小化。l^0 范数最小化是非凸集问题，通常将其转化成 l^1 范数最小化。因此矩阵 $\boldsymbol{F}$ 的稀疏最大化问题可抽象为如下秩最小问题：

$$\begin{aligned} &\text{minimize} \quad \operatorname{rank}(\boldsymbol{F}) \\ &\text{subject to} \quad K(\boldsymbol{F}) = \boldsymbol{Q} \end{aligned} \tag{6}$$

这里并未直接计算该矩阵秩最小化问题，而是设计了一种启发式贪婪算法。算法输入参数为特征全集 F，偏离函数 $V(?)$以及门限 *threshold*；输出受异常影响较大的特征集 A_f。首先初始化所有流特征均未受到异常的影响，即初始化 $N_f=F$，$A_f=\varnothing$,遍历 N_f集合，每次遍历计算 $V(N_f)$和 $V(N_f\backslash f_i)$，如果二者之差异大于门限 *threshold*，那么算法就认为f_i受异常影响较大，将其添加到 A_f，同时从 N_f中剔除。如果 $A_f=\varnothing$，那么 *threshold* 设置的偏大，缩小 *threshold* 后，递归调用；如果集合 F 内特征数与 A_f中之比小于 10，则违背了定理 1 中稀疏化的现象，递归调用，直到满足定理 1 为止。另外，当发现 A_f 仍然违背定理 1 时，为了加快算法的收敛，下次递归前将集合 F 初始化为当前得到的 A_f，以达到逐渐减小集合 F 内所包含特征的目的。

图 2 给出该算法的伪码描述，算法中的 N_f为候选特征集。对 RRF 算法的 2 点说明如下。

1) 算法中用于计算流特征因异常而偏移的函数 $V(?)$的选择有很多，如 PCA、kalman 等，这些算法虽然性能良好，但计算复杂度较高。实验表明选择简单的均值偏差函数 $V(f_i)=|f_i-\mathrm{E}(f_i)|/s_{f_i}$，即偏离均值多少个标准差，即可达到很好的效果。

2) *threshold* 是自适应调整的，初始值理论上可以为任意值，算法在递归过程中会自动调整，但是考虑到收敛速度，根据 $V(?)$的不同而取不同的区间。由于 *threshold* 需要一个学习过程，因此算法起始阶段的收敛需要一定的时间，但是一旦 *threshold* 收敛，算法可以快速地完成筛选。

```
Procedure I      RRF(F, N_last, threshold, V(•))
Input:
    F: the full set of flow features
    N_last: N_f that acquire by last recursion,  initially
    Every recursion makes N_f = N_last to be the full
    set of flow features
    threshold: deviation threshold
    V(•): the function computing the deviations of
    features caused by anomaly
Output:
    A_f: the subset of flow features contained anomaly
    N_f ← F = {f_i | f_i ∈ F};
    A_f ← Φ;
    for all f_i ∈ N_f     do
        if |V(N_f) − V(N_f / f_i)| > threshold then
            A_f = add(A_f, f_i);
            N_f = eliminate (N_f, f_i);
        else
            continue;
        endif
    endfor
    N_last = add(N_last, N_f);
    if A_f = Φ          then
        threshold = reduce(threshold);
        A_f = RRF(F, N_last, threshold, V(•));
    elseif element (A_f ∪ N_last)/ element(A_f) < 10 then
        threshold = augment(threshold);
        F ← A_f;
        A_f = RRF(F, N_last, threshold, V(•));
    endif
    return A_f;
```

图 2　RRF 算法伪码描述

3.2　聚焦分类算法(FCA)

FCA 的思想是流特征f_i的状态的改变是由某异常子集 C_i 中的一个或多个异常引起的；同时单个异

常的发生会引起多个流特征的状态的改变(如表 1 所示)；不同流特征对应着不同的异常子集，多维流特征同时异常时将对应着多个异常子集，其中最频繁出现的异常类将被确认为最终异常。对算法探讨之前给出以下标识符及概念。

C：为异常类全集。

C_i：为特征f_i对应的异常子集。

定义 2 （项）标识异常的每一维流特征f_i称为项，例如 H(*dest_port*)。

定义 3 （频繁集）由 l 项检测某一异常时，l 项对应的 l 个异常集中的异常类出现的次数超过一定阈值时，称这些异常类组成的集合为频繁集。频繁集中的元素称为频繁项，出现次数最多的异常类称为最频繁项。

定义 4 （m-频繁项）由 l 项检测某一异常时，其对应的 l 个异常子集 $C_i(i=1,2,\cdots,l)$中该异常出现 $m(m\leq l)$次，则称该异常类为 m-频繁项，m 为该异常类的支持度。

图 3 给出了 FCA 的伪码描述，其具体步骤如下。

```
Procedure 2.  FCA(F_sub, C^last)
Input:
    F_sub: the subset of features {f_1, f_2, ..., f_r} selected by RRF
    C_last: C_r which latest Output of the FCA, and C initially.
Output:
    C_r: the real class that the anomaly affiliated
    F_sub ← C:{ C_1, C_2, ..., C_l };
    C_r ← C_1 ∩ C_2 ∩ ... ∩ C_l;
    C_r = intersection(C_r, C^last);
    If C_r = Φ   Then
        //计算C中最频繁集
        C_r = max_support(C),
        intersection(C_r, C^last);
    Endif
    Return C_r;

    Procedure 3. intersection(C_r, C^last)
    If elements(C_r) >1 && C^last ≠ Φ  Then
        //求与前一时刻的分类结果集C^last的交集
        C_r = C_r ∩ C^last;
    Endif
    Return C_r;
```

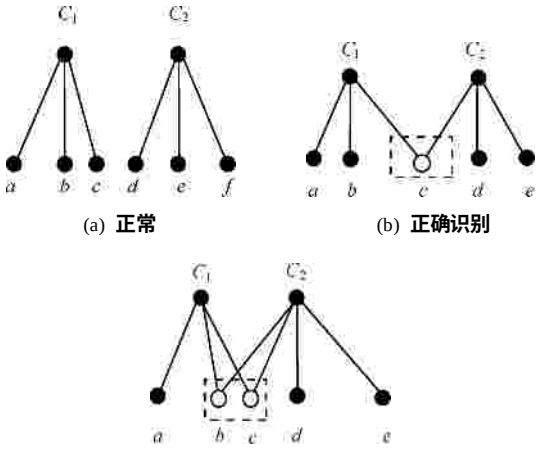
图 3 FCA 算法伪码描述

步骤 1 根据 F_{sub} 映射到模式库中对应的异常集 C，计算 C 中元素的相交和 C_r，即求 C 中的 l-频繁项。

步骤 2 如果 C_r 中的元素个数大于 1，分类出现不确定，将 C_r 与前一时刻得到的异常集 C^{last} 相交。

步骤 3 如果 C_r 不包含任何元素，那么计算 C 的最频繁集。

图 4 给出了分类的 3 种可能的输出，设 a~f 是已知异常类，C_1,C_2 分别是特征 f_1，f_2 引起的异常集，第 1 种情况如图 4(a)所示，$C_r=C_1\cap C_2=\varnothing$，表示没有异常发生；第 2 种情况如图(b)展示了能够正确分类的情况，$C_r=C_1\cap C_2=\{c\}$；第 3 种如图 4(c)展示了一个不确定的情况，$C_r=C_1\cap C_2=\{b, c\}$，考虑到异常的局部聚集特性，将前一时刻的输出反馈回来，以降低不确定性。4.2 节将仿真评估上述情况出现的比率。



(a) 正常 (b) 正确识别

(c) 歧义

图 4 FCA 算法分类可能的结果

3.3 MRLR 模型的普适性推广

MRLR 模型适用于任何分类算法而不局限于某一种。基于 MRLR 模型的 RRF 算法能够动态地从高维流特征中选择与所检测异常强相关的低维流特征。为了验证 MRLR 模型的通用性，本节探讨 MRLR 模型用于另外几种典型的分类算法——基于贝叶斯、基于支持向量机(SVM, support vector machine)和 PCA+Entropy。考虑以下应用场景：设 n 个 c 类 d 维训练样本，分别进行训练得到 c 个模式，根据得到的模式检测异常。

1) MRLR+Bayes。基于贝叶斯的分类算法关键是估计特征的后验概率密度函数。如果贝叶斯参数估计采用增量算法，分类上述场景的计算复杂度为 $O(cnd^2)$根据贝叶斯参数估计。MRLR 模型用于 Bayes 算法的特征选择，去除不相关的特征(减小 d)，保留较低维的强相关流特征，可降低算法复杂度。

2) MRLR+SVM。SVM 中的核函数可以将特征从低维非线性空间映射到高维线性空间，并在此空间做分类决策。SVM 同样面临着特征选择问题。SVM 分类上述场景的计算时间复杂度为 $O(cdn^2)$。将 MRLR 模型用于 SVM 分类，为 SVM 动态选择

“合适的”流特征并将复杂度降低为 $O(cn^2)$。

3) PCA+Entropy。由 PCA 挖掘高维特征中的信息，然后计算各特征分布的熵作为分类参数[5,6]。上述场景的计算时间复杂度为 $O(d^3+d^2n)$。

4) MRLR+FCA。算法简单实用，计算复杂度为 $O(n)$。

4　仿真实验

实验中采用手工标记异常和注入异常 2 类数据集。第一，检验基于 MRLR 模型的 RRF 算法对特征维数的缩减能力及其选择的准确性；第二，测试 MRLR+FCA 的异常分类算法的能力；第三，MRLR 模型用于其它异常分类算法后对其影响；最后，对各种分类算法的时间复杂度进行了比较。

4.1　数据集

试验中使用的数据集如表 2 所示。A 是从某大型企业网采集的流量，由于该企业网采用隧道协议传输，收集的流量比较干净，适合作为背景流量可控地注入异常。数据集 B 和 C 均含有异常且通过人工标记出了 B、C 中的异常。采集数据集 A 和 B 时统计了如表 3 中的特征集，C 的原始记录是流量矩阵，通过汇总处理得到表 3 中的特征集。

表 2　　实验数据

标号	来源	收集时间	收集间隔/min	异常比(异常节点/总节点)
A	企业网	2010-08	1	<1%
B	教育网	2010-04	1	<19%
C	Abilene	2007-05	5	<22%

为了分析 B 和 C 的异常比，分别取一月记录，每小时统计一次，然后一天为单位求统计平均。图 5 给出了分析结果，可以看出：B 中的异常比最高达到了 19%，C 则达到了 22%，但两者分布时段不同。这是因为网络用户不同。教育网中的数据主要来自学生，上网主要集中在课余时间。而 Abilene 网络主要传输实验和研究数据。同时图 5 的结果还验证了关于实际网络中的异常是稀疏分布的结论。

4.2　实验过程

初始流特征集选用如表 3 所示的。针对每一特征序列 $f_i(t)$ 进行 m 层 EMD 多尺度分解得到流特征矩阵 $\boldsymbol{F}(i,j)$ $(i=1,2,\cdots,N;\ j=1,2,\cdots,M)$，其中每一行代表一个尺度序列，每一列表示某一时刻(某一时刻事实上表示设定的统计时间段)的各个尺度分量的值。若初始特征集中特征维数为 n，那么 $N=mn$。例如，本实验 $n=11$，m 的值是自适应的，一般取 $m=7\sim10$，那么特征维数也将在 77~110 之间。在进行实验前，首先定义以下概念。

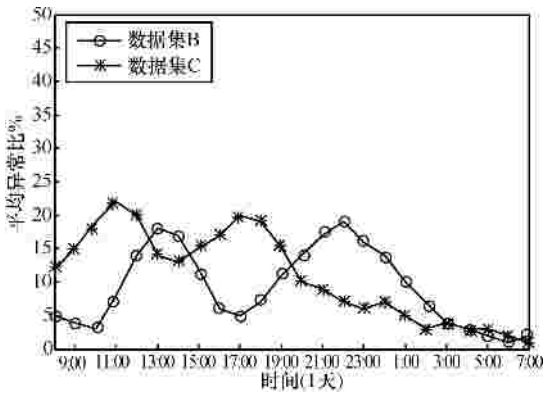


图 5　数据集 B 和 C 的异常比变化趋势(1 天)

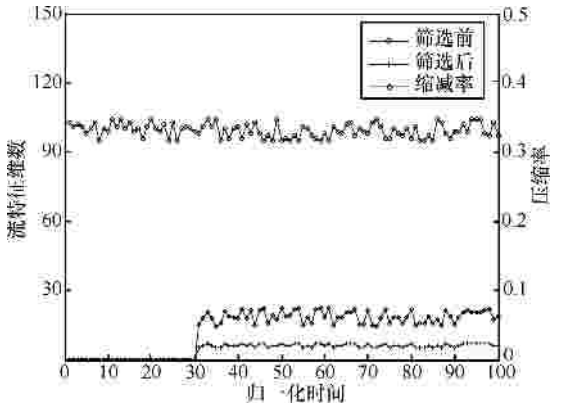
表 3　　初始特征集

标号	描述
f_1	由五元组标识的不重复流个数
f_2	分组数量
f_3	字节数
f_4	不重复源 IP 总数
f_5	不重复目的 IP 总数
f_6	不重复源 port 总数
f_7	不重复目的 port 总数
f_8	源 IP 的信息熵
f_9	目的 IP 的信息熵
f_{10}	源端口的信息熵
f_{11}	目的端口的信息熵

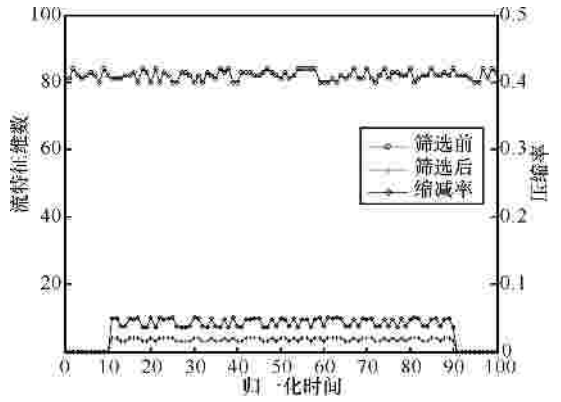
定义 5　（缩减率）设初始流特征集的特征维数为 n，由 RRF 算法识别的异常流特征维数为 l，称 l/n 为缩减率。

定义 6　（异常密度）设某网络总的链路数为 m，任意时刻该网络中出现异常的链路数为 k，称 k/m 为异常密度。

实验采用数据集 A 作为背景流量，可控地注入异常，观察 RRF 算法前后输入空间的维数变化。图 6(a)，图 6(b)分别给出了注入 DDoS 和 Network Scan 后 RRF 标识的异常流特征和输入流特征维数以及特征缩减率。可以看出：RRF 算法对流特征维数的缩减率基本在 10%以下，因此其极大降低了特征维数。



(a) DDoS



(b) Network Scan

图 6 RRF 算法缩减输入特征维数仿真结果

Neyman-Pearson 理论中关于统计检验问题定义了 FP(false positive rate)与 TP(true positive rate)，当门限较大时 TP 较大，FP 也会增大，反之亦然。实验中采用综合了两者的 ROC(receiver operating characteristics)曲线检验算法性能。ROC 曲线将 TP 表示为 FP 的函数，因此左上角的曲线性能最好。图 7 给出了分别注入 DDoS 和 Network Scan 异常后 RRF 算法 ROC 曲线。可以看出：RRF 能够以较低的误报率达到较高的检测率。

图 8 给出了 RRF 算法在不同异常密度 r 下的 ROC 曲线。可以看出：RRF 算法随着异常密度的增加，其性能不断地下降。当异常密度小于 30%时，RRF 算法在低误报率的情况下有较高检测率。当异常密度大于 30%后，算法性能急剧恶化。这是因为 MRLR 模型成立的前述假设失效，算法性能急剧下降。

图 9 采用数据集 B 给出了 3.2 节讨论的 FCA 算法的 3 种结果出现的比例。可以看出，不确定(ambiguous)成分所占的比例很小，而正确分类超过 90%(normal+correct)。这是因为数据集 B 的异常密度较小，异常之间出现交叉的现象很少。

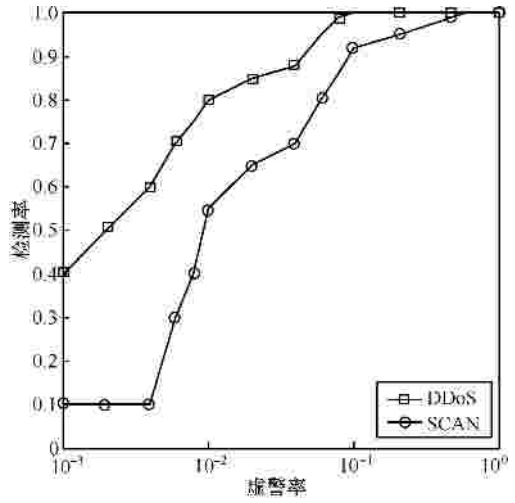


图 7 RRF 算法对 2 种异常情况下的 ROC 曲线

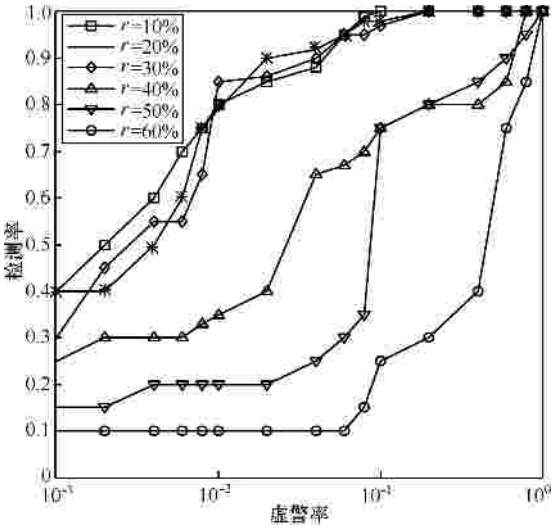


图 8 不同异常密度下 RRF 算法的 ROC 曲线

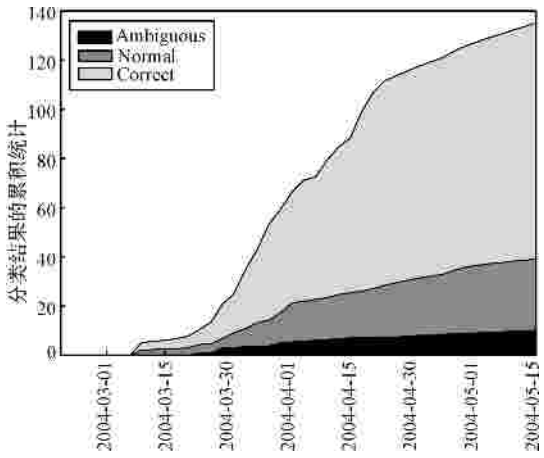
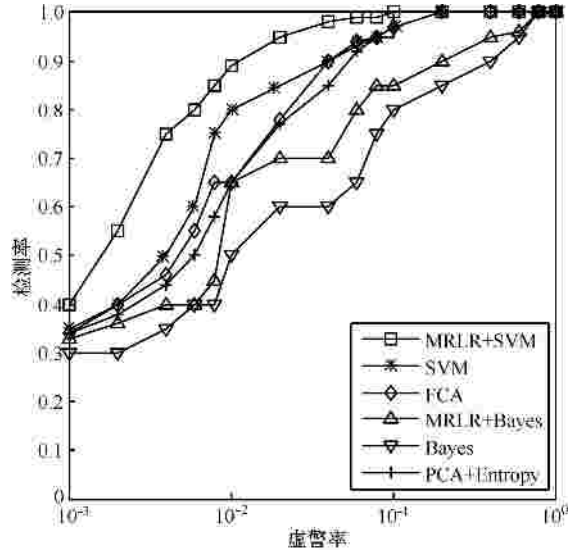


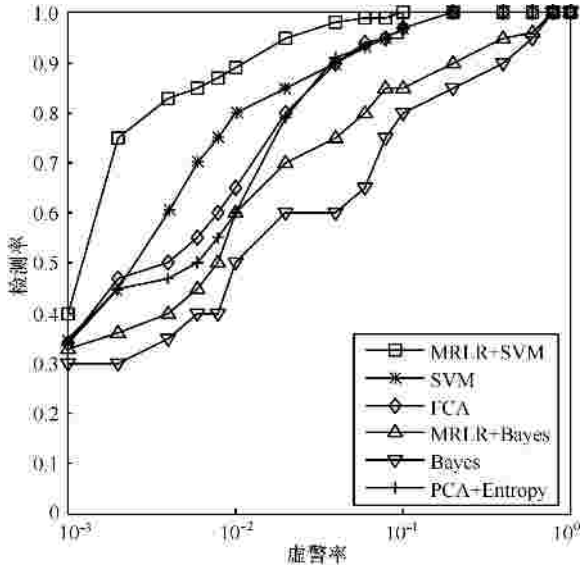
图 9 FCA 分类的性能

图 10 给出了在异常密度小于 0.3 时，FCA 算法与 SVM、Bayes、MRLR+SVM、MRLR+Bayes 和 PCA+Entropy 的性能比较。可以看出：SVM,

Bayes 与 MRLR 相结合后性能有较大改善。FCA 的分类精度低于 MRLR+SVM 和 SVM，优于 MRLR+Bayes、Bayes 算法，略优于 PCA+Entropy。但是，FCA 的计算复杂度仅为 $O(n)$(n 为输入空间的维数)，远低于其余算法。



(a) 采用数据集 B



(b) 采用数据集 C

图 10　基于 MRLR 模型的多种分类算法的性能比较

最后表 4 给出了 FCA 与其他算法的时间复杂度的比较，可以看出：同等条件下 MRLR+FCA 的算法复杂度远低于 SVM，Bayes 和 PCA+ Entropy 算法；另外，经过 MRLR 模型缩减特征维数后，Bayesian 和 SVM 的耗时分别下降了约 25% 和 30%。

表 4　　算法时间复杂度比较

算法	数据集 B	数据集 C
Bayesian	95	93
SVM	102	98
MRLR+Bayesian	69	72
MRLR+SVM	72	68
MRLR+FCA	29	43
PCA+entropy	75	97

5　结束语

本文根据网络异常的稀疏分布特性提出了 MRLR 模型描述网络中异常的分布，并基于此模型设计了一种特征动态选择(RRF)算法，并将 RRF 算法的输出的低维特征子集作为输入，提出了一种简单的分类算法(FCA)。实验结果表明了模型的有效性。

参考文献:

[1] BARFORD P, KLINE J, PLONKA D. A signal analysis of network traffic anomalies[A]. Proceedings of IMW[C]. Marseille, 2002.71-82.

[2] SOULE A, SALAMATIAN K, TAFT N. Combining filtering and statistical methods for anomaly detection[A]. Proceedings of IMC[C]. Berkeley, CA, USA, 2005. 331-344.

[3] SILVEIRA F, DIOT C, TAFT N. ASTUTE: Detecting a Different Class of Traffic Anomalies (Extended Version)[R]. Technicolor, 2010.

[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.
ZHUGE J W, WANG D W, CHEN Y, *et al*. A network anomaly detector based on the D-S evidence theory[J]. Journal of Software, 2006, 17(3):463-471.

[5] LAKHINA A, CROVELLA M, DIOT C. Mining anomalies using traffic feature distributions[A]. Proceedings of SIGCOMM[C]. Philadelphia, PA, USA, 2005. 217-228.

[6] LAKHINA A, CROVELLA M, DIOT C. Diagnosing network-wide traffic anomalies[A]. Proceedings of SIGCOMM[C]. Portland, Oregon, USA, 2004. 219-230.

[7] RINGBERG H, SOULE A, REXFORD J. Sensitivity of PCA for traffic anomaly detection[A]. Proceedings of ACM SIGMETRICS'07[C]. San Diego, 2007.

[8] 徐琴珍, 杨绿溪. 一种基于有监督局部决策分层支持向量机的异常检测方法[J]. 电子与信息学报, 2010, 32(10), 2383-2387.
XU Q Z, YANG L X. A supervised local decision hierachical support vector machine based anomaly intrusion detection method[J]. Journal of Electronics & Information Technology, 2010, 32(10), 2383-2387.

[9] LELAND W E, TAQQU M S, WILLINGER W. On the self-similar nature

of ethernet traffic[J]. Transactions on Networking, 1994, 2(1):1-15.

[10] PAXSON V, FLOYD S. Wide-area traffic: the failure of poisson modeling[J]. IEEE/ACM Transactions on Networking, 1995,1(3): 226-244.

[11] GARCIA R C, SADIKU M N O, CANNADY J D. WAID: wavelet analysis intrusion detection, circuits and systems[A]. The 2002 45th Midwest Symposium[C]. Tulsa, 2002. 688-691.

[12] **任勋益**, **王汝传**, **王海艳**. **基于自相似检测** DDoS **攻击的小波分析方法**[J]. **通信学报**, 2006, 27(5): 6-11.
REN X Y, WANG R C, WANG H Y. Wavelet analysis method for detection of DDoS attack based on self-similar[J]. Journal on Communications, 2006, 27(5): 6-11.

[13] **许晓东**, **朱士瑞**, **孙亚民**. **基于分形特性的宏观网络流量异常分析**[J]. **通信学报**, 2009, 30(9): 43-53.
XU X D, ZHU S R, SUN Y M. Anomaly detection algorithm based on fractal characteristics of large-scale network traffic[J]. Journal on Communications, 2009, 30(9): 43-53.

[14] SILVEIRA F, DIOT C, TAFT N. ASTUTE: detecting a different class of traffic anomalies[A]. Proceedings of SIGCOMM[C]. New delhi, India, 2010.

[15] ZHANG Y, ROUGHAN M, WILLINGER W. Spatio-temporal compressive sensing and internet traffic matrices[A]. Proceedings of SIGCOMM[C]. Barcelona, Spain, 2009.

作者简介：

程国振（1986-），**男，山东定陶人，国家数字交换系统工程技术研究中心博士生，主要研究方向为网络安全和异常流量检测。**

程东年（1957-），**男，河南原阳人，国家数字交换系统工程技术研究中心教授，主要研究方向为宽带信息网络体系结构、网络安全协议和网络性能分析技术。**

俞定玖（1965-），**男，四川什邡人，国家数字交换系统工程技术研究中心副教授，主要研究方向为移动通信。**