DOI:10.13718/j.cnki.xsxb.2020.09.021

# 一种采用遗传规划优化的基于非负矩阵分解的入侵检测模型①

吴 渝[1], 彭茂玲[2], 钱仁飞[3], 马哲峰[3], 陈善雄[4]

1. 重庆警察学院, 重庆 401331; 2. 重庆城市管理职业学院, 重庆 401331;
3. 宁波大发化纤有限公司, 浙江 宁波 315336; 4. 西南大学 计算机与信息科学学院, 重庆 400715

**摘要**: 利用非负矩阵分解对系统监控的特权程序行为进行检测, 进而发现异常的入侵行为. 在矩阵分解过程中采用 Alpha 散度作为度量标准, 并引入遗传规划算法对非负矩阵分解中的误差函数进行优化. 实验中, 对新墨西哥大学的 Sun SPARC 工作站上获取的系统调用数据和 CICIDS2017 数据集进行了测试, 测试结果表明, 该方法相对于传统的检测方法而言, 在入侵检测精度方面具有良好性能.

**关 键 词**: 入侵检测; 系统调用; 非负矩阵分解; 遗传规划

**中图分类号**: TP393.0 **文献标志码**: A **文章编号**: 1000-5471(2020)09-0139-08

入侵检测可以在恶意攻击发生时或发生后直接检测到它们. 它可以从网络或计算机系统的某个节点收集流量或系统调用等信息, 分析和判别数据特征, 进而区分正常和异常行为, 从而实现对网络的保护[1-2]. 一般来说, 入侵检测系统需要处理大量的数据, 这些数据可能包含一些不相关和冗余的特性, 这些特性会减慢训练和测试过程, 增加资源消耗, 降低检测率. 特征选择是入侵检测系统的核心内容之一[3], 从网络数据中提取正常或异常特征是自动生成入侵特征的最有效方法之一. 如果特征知识能够被恰当地提取和表达出来, 还可以预测新的攻击[4]. 特征提取可以有效地降低计算维数, 加快入侵检测进程[5]. 本文提出了一种通过监控特权程序行为来检测入侵行为的方法, 为了建立异常检测模型, 考虑了系统调用的特征, 然后, 这些系统调用被按照一定长度进行序列化, 以生成数据向量来训练分类器. 在此基础上, 利用非负矩阵分解(nonnegative matrix factorization, NMF)模型来分析特权程序的正常和异常行为, 结合 Alpha 散度来度量迭代分解过程中的差异程度, 并利用遗传算法优化非负矩阵分解中的误差函数, 以此来提升检测的精度.

## 1 基于非负矩阵分解的入侵检测

非负矩阵分解方法的一个优点是计算效率较高[6-7]. 基于 NMF 的模型对系统调用的分析可以避免对单个程序调用序列进行训练, 降低了学习过程的计算成本[8-9]. 本文利用系统调用序列中的特征向量构建 NMF 模型的输入矩阵. 通过矩阵非负分解得到基矩阵, 以此来创建程序的行为特征, 从而构建起特征库. 计算需要检测的测试向量与其重建矩阵之间的距离, 并与阈值进行比较, 以确定是否存在异常行为.

### 1.1 非负矩阵分解

在文献[10]中引入了非负矩阵因式分解方法, 寻找非负部分的数据表示. 给定一个非负矩阵 $\boldsymbol{V}$, 寻找非负矩阵因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$, 使得

① 收稿日期: 2019-11-16
基金项目: 国家自然科学基金项目(61303227); 重庆市自然科学基金项目(cstc2020jcyj-msxm2695); 中国博士后基金项目(2015M580765); 重庆市博士后科研项目(Xm2016041); 中央高校基本科研业务费项目(XDJK2018B020).
作者简介: 吴 渝(1980-), 男, 讲师, 工程师, 主要从事计算机软件开发研究, 网络信息安全管理和信息化建设管理.
通信作者: 彭茂玲, 教授.

$$\boldsymbol{V} \approx \boldsymbol{WH} \tag{1}$$

其中 $\boldsymbol{W}$ 是一个 $m \times r$ 的矩阵而 $\boldsymbol{H}$ 是一个 $r \times n$ 的矩阵. (1) 式也可以按照列相乘的方式写为 $\boldsymbol{v}_i \approx \boldsymbol{W}\boldsymbol{h}_i$, $\boldsymbol{v}_i$ 和 $\boldsymbol{h}_i$ 分别表示 $\boldsymbol{V}$ 的第 $i$ 列和 $\boldsymbol{H}$ 的第 $i$ 列向量, 也即矢量 $\boldsymbol{v}_i$ 是 $\boldsymbol{W}$ 的列与权重 $\boldsymbol{h}_i$ 的线性组合. $\boldsymbol{W}$ 的列可以表示基向量, 且当与适当的混合权重($\boldsymbol{H}$ 的某列) 组合时, 将得到 $\boldsymbol{V}$ 的线性近似值.

### 1.2 基于 Alpha 散度的误差度量

非负矩阵在分解过程中通过不断迭代产生近似的基矩阵 $\boldsymbol{W}$ 和权矩阵 $\boldsymbol{H}$, 其分解过程中度量差异程度的指标采用 Kullback-Leibler(KL) 散度[11]. KL 散度常用于计算两个概率分布 $P$ 和 $Q$ 的差异, 表示当用概率分布 $Q$ 来拟合真实分布 $P$ 时产生的信息损耗, 在非负矩阵分解中用来度量 $\boldsymbol{V}$ 与 $\boldsymbol{WH}$ 的差异. 但 KL 散度无法对分解过程进行稀疏约束, 而通常系统调用数据很大程度上是稀疏数据, 导致分解过程有可能收敛到局部最优值, 因此引入 Alpha 散度作为度量标准. Alpha 散度属于凸散度集, 该散度集属 csiszár 散度集, 也称 Ali-silvey 散度[12-13]. 所以本文考虑基于 Alpha 散度的 NMF 的目标函数, 它被用于度量 $\boldsymbol{V}$ 和 $\boldsymbol{WH}$ 之间的差异, 其定义为

$$D_\alpha[\boldsymbol{V} \,||\, \boldsymbol{WH}] = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^{m}\sum_{j=1}^{n} \alpha v_{ij} + (1-\alpha)[\boldsymbol{WH}]_{ij} - v_{ij}^{\alpha}[\boldsymbol{WH}]_{ij}^{1-\alpha} \tag{2}$$

(2) 式的 Alpha 散度通常可以表达为

$$D_A^\beta(\boldsymbol{WH} \,||\, \boldsymbol{V}) = \sum_{i=1}^{m}\sum_{k=1}^{n} v_{ik} \frac{\left(\frac{v_{ik}}{z_{ik}}\right)^{\beta-1} - 1}{\beta(\beta-1)} + \frac{z_{ik} - v_{ik}}{\beta} \tag{3}$$

$$z_{ik} = [\boldsymbol{WH}]_{ik} \tag{4}$$

其中 $\beta = \frac{1+\alpha}{2}$. 事实上, 我们可以通过多种方式对 Alpha 散度进行平滑或稀疏约束, 以增强 Alpha 散度的数据特征表达能力. 本文引入了一个带正则项的修正 Alpha 散度

$$D_A^\beta(\boldsymbol{WH} \,||\, \boldsymbol{V}) = \sum_{i=1}^{m}\sum_{k=1}^{n} v_{ik} \frac{\left(\frac{v_{ik}}{[\boldsymbol{WH}]_{ik}}\right)^{\beta-1} - 1}{\beta(\beta-1)} + \frac{[\boldsymbol{WH}]_{ik} - v_{ik}}{\beta} + \alpha_s J(\boldsymbol{H}) + \alpha_s J(\boldsymbol{W}) \tag{5}$$

其中: $\alpha_s \geqslant 0$ 是正则化参数; 引入项 $\boldsymbol{J}(\boldsymbol{H})$ 和 $\boldsymbol{J}(\boldsymbol{W})$ 用来强化依赖于应用的特性, 如光滑性或稀疏性[14], 这里

$$\boldsymbol{J}(\boldsymbol{X}) = \sum_{j=1}^{m}\sum_{k=1}^{n} f_X(x_{jk}) = \sum_{j=1}^{m}\sum_{k=1}^{n} x_{jk} \tag{6}$$

为了得到稀疏表达, 通常选择 $f(x_j) = x_j$ 或者 $f(x_j) = x_j \log(x_j)$, $x_j \geqslant 0$, 因此, $\boldsymbol{J}(\boldsymbol{H})$ 和 $\boldsymbol{J}(\boldsymbol{W})$ 沿用了式 (6) 定义. 对(5) 式应用梯度下降方法, 得到非负矩阵分解的乘性迭代规则

$$h_{jk} \leftarrow \left(h_{jk}\left(\sum_{i=1}^{m} w_{ij}\left(\frac{v_{ik}}{[\boldsymbol{WH}]_{ik}}\right)^{\beta}\right)^{\omega\cdot\beta^{-1}}\right)^{1+\alpha_s} \tag{7}$$

$$w_{ij} \leftarrow \left(w_{ij}\left(\sum_{k=1}^{n} h_{jk}\left(\frac{v_{ik}}{[\boldsymbol{WH}]_{ik}}\right)^{\beta}\right)^{\omega\cdot\beta^{-1}}\right)^{1+\alpha_s} \tag{8}$$

其中: $w_{ij} \geqslant 0$, $h_{jk} \geqslant 0$, $\forall\, i,j,k \in \mathbb{N}$. 在每一步迭代中 $w_{ij}$ 将被正则化:

$$w_{ij} \leftarrow \frac{w_{ij}}{\left(\sum_{p} w_{pj}\right)} \tag{9}$$

这里 $\omega \in (0, 2)$ 是过松弛参数. 本文应用非线性变换 $x_{jk} = (x_{jk})^{1+\alpha_s}$, 其中 $\alpha_s$ 是一个取值较小的系数, 通常取值在 0.001 与 0.005 之间. 如果想增加计算过程中的稀疏性, 取正值; 如果想减少稀疏性, 取负值. 类似于 Alpha 散度, 还设计了 Jensen-Shannon 散度乘法迭代规则($\beta = 2$), Hellinger 散度的乘法迭代规则($\beta = 0.5$), 这些规则在稀疏性和松弛性方面施加了额外的约束.

### 1.3 基于非负矩阵分解的入侵检测

本文以系统命令调用序列和系统访问记录为数据源, 对入侵行为进行分析. 利用非负矩阵分解方法建立系统调用行为的检测及分析模型. 通过方程(7),(8),(9) 给出的规则把 $\boldsymbol{V}$ 分解为两个矩阵 $\boldsymbol{W}_{\text{training}}$ 和 $\boldsymbol{H}$. 矩阵 $\boldsymbol{W}$ 训练的每一列都包含一个基向量, 而 $\boldsymbol{H}$ 的每一列表示 $\boldsymbol{V}$ 中对应列所需的系数.

事实上，对于等式 $\boldsymbol{v}_i \approx \boldsymbol{W}\boldsymbol{h}_i$，$\boldsymbol{v}$ 和 $\boldsymbol{h}$ 分别对应 $\boldsymbol{V}$ 和 $\boldsymbol{H}$ 的列向量. 每个向量 $\boldsymbol{v}$ 由 $w$ 列的线性组合来逼近，而 $w$ 列的加强分量为 $\boldsymbol{h}$. 基于此，对于测试矩阵 $\boldsymbol{V}_{\text{test}}$（由测试数据构成）的列向量而言，使用从规一化训练数据中得到的基 $\boldsymbol{W}_{\text{training}}$，通过迭代更新规则(10) 能够找到编码系数的表示向量 $\boldsymbol{h}$，然后重构 $\boldsymbol{v}$ 得到 $\boldsymbol{v}'$

$$\boldsymbol{v}' = \boldsymbol{W}_{\text{training}} \times \boldsymbol{h} \tag{10}$$

然后利用式(5) 计算两个向量之间的相似性

$$\lambda = \boldsymbol{D}(\boldsymbol{v} \mid \boldsymbol{v}') \tag{11}$$

如果测试数据是归一化的，则通过式(10) 重构的 $\boldsymbol{v}'$ 与 $\boldsymbol{v}$ 非常相似，$\lambda$ 较小. 测试数据向量是否正常可以通过 $\lambda$ 与阈值的比较来进行判别. 当 $\lambda < \varepsilon$ 时为正常行为，否则分类为异常行为，其中 $\varepsilon$ 为阈值，这是一个二分类问题. 整个入侵检测的判断过程被分为如下三个步骤：

步骤 1：在训练阶段找到用 Ahpha 散度来度量 NMF 的最佳参数，如 $\beta,\omega,\alpha_s$，然后将寻优问题转化为多峰函数优化过程，进一步针对入侵检测模型，应用遗传规划算法来获取 $\beta,\omega,\alpha_s$ 的最佳值.

步骤 2：在最佳的 $\beta,\omega,\alpha_s$ 参数状态下，用训练数据建立基于 NMF 的学习模型. 利用训练集构建矩阵 $\boldsymbol{V}$，通过式(7)，(8) 的矩阵分解得到 $\boldsymbol{W}_{\text{training}}$ 及对应系数 $\boldsymbol{H}$.

步骤 3：利用相应的乘法学习规则对测试数据进行分解，然后进行重构，利用式(11) 对入侵过程和正常过程进行分类.

## 2　利用遗传规划对基于 NMF 的入侵检测方法进行优化

遗传算法是一类基于进化和自然选择原理的计算模型. 这些算法通过使用类似染色体的数据结构将特定域中的问题转换为模型，并使用选择、重组和变异算子进化染色体[14-15]. 本文利用遗传算法分别得到非负矩阵分解中的 $\beta,\omega,\alpha_s$ 等参数，也即利用遗传算法学习和优化非负矩阵分解的入侵检测模型. 具体的实现过程如下：

1) 读入非负矩阵分解迭代方程；

2) 读入任务参数$\{\beta,\omega,\alpha\}$；

3) 读入 GP 参数：种群规模($p_S$)、最大进化代数($G_{\max}$)、遗传算子 GO 集合、遗传算子概率因子向量 $\boldsymbol{P}_{GO}$；

4) 根据$\{\beta,\omega,\alpha\}$ 构造种群规模为 $p_S$ 的初始群体；

5) 计算每个个体染色体的适应度函数，保存最好的个体；旋转赌轮选择 $p_S$ 个个体作为父代；对父代按照遗传算子执行进化操作以产生新一代群体；将新一代群体代替当前群体；重复上述过程，直至进化代数达到 $G_{\max}$；

6) 返回最好个体的染色体以得到非负矩阵分解中的最佳参数 $\beta,\omega,\alpha_s$.

该算法核心是将遗传规划方法应用于基于非负矩阵分解迭代方程式中，其中每个个体的染色体用变长的层次结构来表示一组参数组合方案. 算法通过对群体的多个个体执行进化操作以促进群体的进化，并利用适应度函数作为启发信息来指导算法朝规划的优化方向发展.

## 3　实验与分析

### 3.1　数据分析

收集了几个由活动进程执行的系统调用数据集. 这些程序的大小和复杂性因不同的入侵方式(缓冲区溢出、符号链接攻击和特洛伊木马程序) 而有很大的不同. 每个跟踪是单个进程从执行开始到结束发出的系统调用的列表. 为了评估所提出的异常检测方法，在实验中使用了两组系统调用数据来分析程序行为.

#### 3.1.1　sendmail 的调用指令集合

实验数据来源于 UNM 集合中关于 sendmail 的调用指令，可以从 http：//www. cs. unm. edu 获得. 跟踪系统调用的指令列表从 UNM 的 Sun SPARC 工作站上得到，该工作站运行的系统为 SunOs4. 1. 1 和 4. 1. 4. 使用 plus 表示正常调用的数据，建立调用长度为 $k = 6$ 的序列长度数据库. 例如，观测到入侵系统调用序列：

Wait4 open　creat　link　creat　write　mkond　open

考虑在SunOS 4.1x操作系统中执行这个任务需要182个系统调用，每一个系统调用被分配的任务号从1到182，系统调用的路径可以表示为数字序列：

7 5 8 9 4 14 5

由于设定序列长度为6，得到3组对应的系统调用序列：

7 5 8 9 4, 5 8 9 4 14, 8 9 4 14 5

实验中，测试数据包含正常数据“bounce”，“bounce-1”，“bounce-2”和“queue”，以及异常数据“sm-280”，“sm-314”，“sm-10763”，“sm-10801”，“sm-10814”. 构建长度为 $k=6$ 的系统调用序列数据库. 利用非负矩阵分解进行异常检测时，在训练阶段从数据库中调用序列构建起矩阵 $\boldsymbol{V}$，通过迭代分解得到矩阵 $\boldsymbol{W}_{\text{training}}$ 和 $\boldsymbol{H}$，检测阶段利用未知序列构建矩阵进行分解得到系数向量 $\boldsymbol{h}$，通过 $\boldsymbol{W}_{\text{training}} \times \boldsymbol{h}$ 重构 $\boldsymbol{v}'$，再与 $\boldsymbol{v}$ 比较就可以判断是否存在异常调用.

3.1.2 CICIDS2017数据集

加拿大网络安全研究所提供了一个名为CICIDS2017的最新数据集，包含最新的威胁和特性. 该数据集吸引全球的研究人员来分析和开发新的模型和算法[16-18]，是当前比较全面的网络安全测试数据集. 该数据集包含8个子集[19].

该数据集拥有11个特征，即匿名性、攻击多样性、完全捕获、完全交互、完全网络配置、可用协议、完全流量、特征集、元数据、异构性和标签[20]. CICIDS2017集合有8个文件，共计2 830 743条记录，每条记录有79个类别. 每一记录行被标记为正常(Normal)或14种攻击类型之一. 表1总结了不同攻击类型和正常行为的分布情况.

**表1 CICIDS2017数据的类别分布**

| 序号 | 数据类型 | 样本量 | 占攻击样本的比例/% | 占总样本的比例/% |
|---|---|---|---|---|
| 1 | Normal | 2 359 087 | / | 83.344 0 |
| 2 | Bot | 1 966 | 0.417 0 | 0.069 4 |
| 3 | DDoS | 41 835 | 8.873 6 | 1.478 0 |
| 4 | DoSGoldenEye | 10 293 | 2.183 3 | 0.363 6 |
| 5 | DoS Hulk | 231 072 | 49.012 7 | 8.163 5 |
| 6 | DoS Slow-httptest | 5 499 | 1.166 4 | 0.194 3 |
| 7 | DoSslowloris | 5 796 | 1.229 4 | 0.204 8 |
| 8 | FTP-Patator | 7 938 | 1.683 7 | 0.280 4 |
| 9 | Heartbleed | 11 | 0.002 3 | 0.000 4 |
| 10 | Infiltration | 36 | 0.007 6 | 0.001 3 |
| 11 | PortScan | 158 930 | 33.710 7 | 5.614 8 |
| 12 | SSH-Patator | 5 897 | 1.250 8 | 0.208 3 |
| 13 | Web Attack Brute Force | 1 507 | 0.319 7 | 0.053 2 |
| 14 | Web AttackSql Injection | 21 | 0.004 5 | 0.000 7 |
| 15 | Web Attack XSS | 652 | 0.138 3 | 0.023 0 |

根据表1中描述的数据分布提取训练和测试子集. 在每个子集中，我们试图包含所有攻击的样本，但是不能在两个子集中出现相同的记录. 提取数据集中每种类型的第一条记录组成训练子集. 在抑制训练子集中随机选择记录组成测试记录.

## 3.2 结果分析

使用召回率 $P_{\text{recall}}$、精确率 $P_{\text{precision}}$、假阳性率 $p_{\text{FPR}}$ 和假阴性率 $p_{\text{FNR}}$ 作为指标来评估算法的性能.

$$P_{\text{recall}} = \frac{C_{\text{TP}}}{C_{\text{TP}} + C_{\text{FN}}}$$

$$P_{\text{precision}} = \frac{C_{\text{TP}}}{C_{\text{TP}} + C_{\text{FP}}}$$

$$p_{\text{FPR}} = \frac{C_{\text{FP}}}{C_{\text{FP}} + C_{\text{TN}}}$$

$$p_{\text{FNR}} = \frac{C_{\text{FN}}}{C_{\text{TP}} + C_{\text{FN}}}$$

其中：$C_{\text{TP}}$ 指正确分类的正样本数，表示根据公式(9)正确判断为异常事件的数量；$C_{\text{FP}}$ 指被错误标记为正

样本的负样本数，表示根据公式(9)判断为异常事件的正常事件数；$C_{TN}$ 指正确分类的负样本数，表示根据公式(9)正确判断为正常事件的数量；$C_{FN}$ 指被错误标记为负样本的正样本数，表示根据式(9)判断为正常事件的异常事件数.

图 1 为对 sendmail 的调用指令集合与 CICIDS2017 数据集的 50 次迭代过程中 $p_{FPR}$ 值和遗传群体平均值的变化曲线. 图 2 为对 sendmail 的调用指令集合与 CICIDS2017 数据集的 50 次迭代过程中 $p_{FNR}$ 和遗传群体平均值的变化曲线. 可以看到在两种数据集下，遗传群体平均值逐渐趋于平稳，说明本文所提算法随着迭代次数的增加而收敛；同时 $p_{FNR}$ 和 $p_{FPR}$ 也稳定到最小值. 显然，$p_{FPR}$ 和 $p_{FNR}$ 的最优值是不同的. 为了找到最合适的参数，对 $p_{FPR}$ 和 $p_{FNR}$ 进行了加权，$p_{FPR}$ 和 $p_{FNR}$ 加权值的变化曲线如图 4 所示. 由图 3 可知，加权最优值为 0.048($p_{FPR}=0.02$，$p_{FNR}=0.028$)，参数 $\beta=2$，$\omega=0.45$，$\alpha_s=0.032$.

图 4 用召回率 $P_{recall}$、精确率 $P_{precision}$ 曲线展示了遗传群体平均值对检测的影响. 可以看出，$p_{FPR}$ 和 $p_{FNR}$ 进行加权后，sendmail 数据集召回率稳定在 0.93，精确率稳定在 0.96；CICIDS2017 数据集召回率稳定在 0.92，精确率稳定在 0.94.

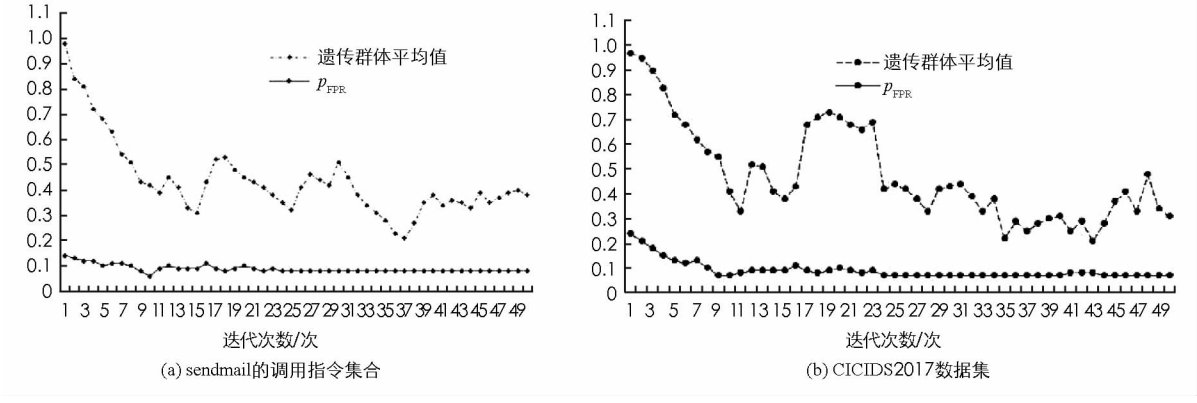



图 1　遗传群体 $p_{FPR}$ 和平均值的变化曲线

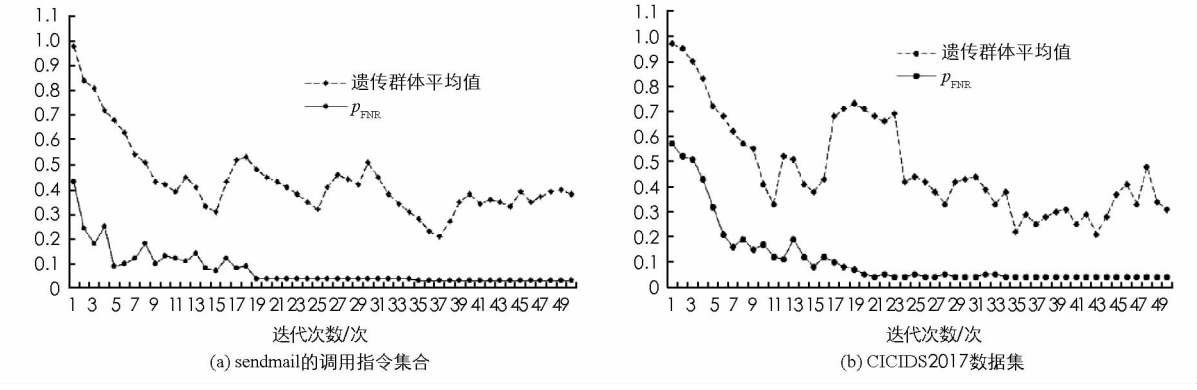



图 2　遗传群体 $p_{FNR}$ 和平均值的变化曲线

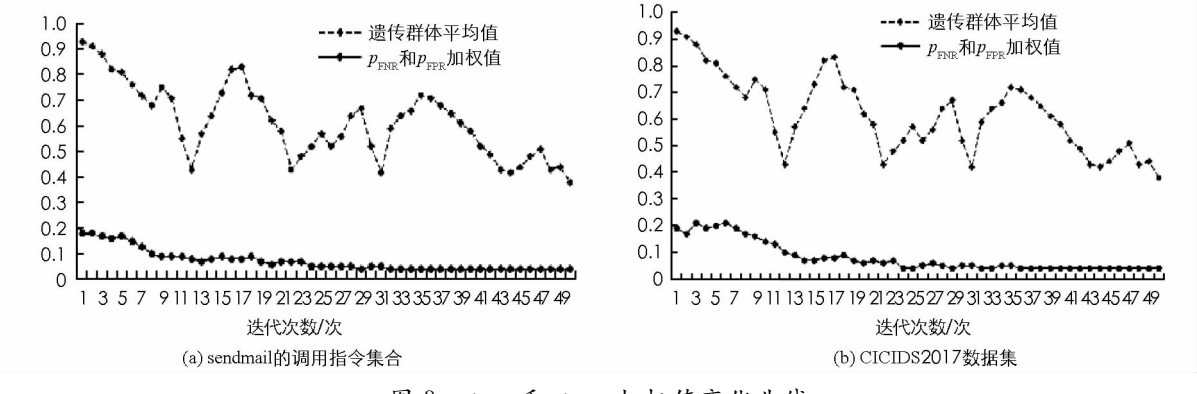
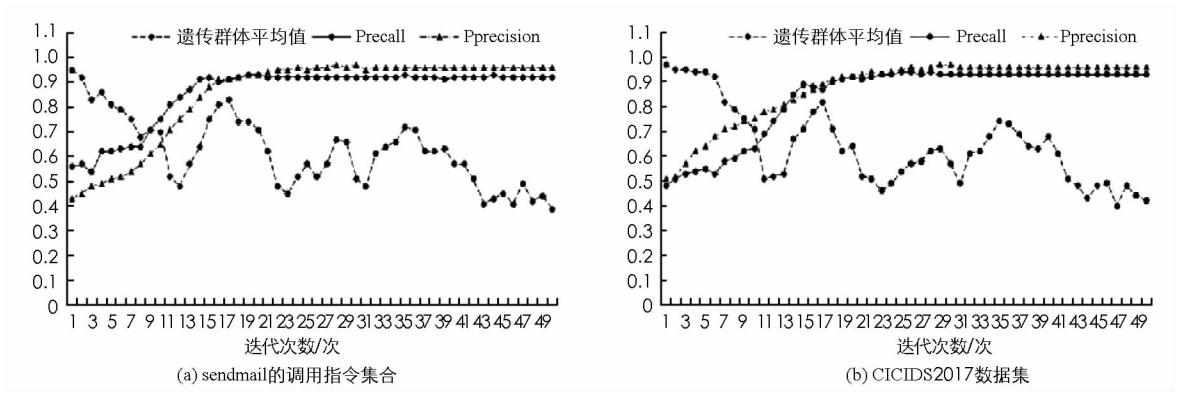



图 3　$p_{FPR}$ 和 $p_{FNR}$ 加权值变化曲线




(a) sendmail的调用指令集合 (b) CICIDS2017数据集

图4 召回率和精确率曲线

为了进一步评估Alpha-NMF模型的参数精度，将Alpha-NMF,PCA,HMM以及标准的NMF进行了对比测试(表2)，其中：异常检测率是指检测出的异常行为数量与样本所有异常行为数量的比率；正常检测率是指检测出的正常行为数量与样本所有正常行为数量的比率.

**表2 Alpha-NMF与其他方法的对比**

| 方法 | sendmail的调用指令集合 | | CICIDS2017数据集 | |
|---|---|---|---|---|
| | 异常检测率/% | 正常检测率/% | 异常检测率/% | 正常检测率/% |
| PCA | 75.3 | 79.0 | 76.3 | 81.2 |
| HMM | 83.3 | 91.3 | 83.7 | 89.5 |
| NMF | 84.4 | 82.0 | 81.3 | 86.2 |
| Alpha-NMF-GP($\beta=2$, $\omega=0.45$, $\alpha_s=0.032$) | 90.5 | 94.3 | 91.2 | 93.3 |
| Alpha-NMF-GP($\beta=0.5$, $\omega=0.3$, $\alpha_s=0.02$) | 89.3 | 90.3 | 88.6 | 90.1 |
| Alpha-NMF-GP($\beta=2$, $\omega=0.26$, $\alpha_s=0.017$) | 86.3 | 88.5 | 87.4 | 88.3 |

从表2可以看出，在sendmail的调用指令集合和CICIDS2017数据集中，Alpha－NMF的异常检测率和正常检测率都高于其他方法. 相对于PCA和HMM方法，本文的方法无论是对异常行为的检测还是正常行为的判断，其性能都有10%左右的提升. 此外，还观察到当$\beta=0.5$，$\omega=0.3$，$\alpha_s=0.02$或$\beta=2$，$\omega=0.26$，$\alpha_s=0.017$时，本文方法性能稍差. 当$\beta=2$，$\omega=0.45$，$\alpha_s=0.032$时，Alpha-NMF能够得到全局最优解，使得本文方法在sendmail集合中达到最佳检测率为90.5%和94.3%. 在CICIDS2017数据集中，最佳检测率为91.2%和93.3%. 同时，实验还对比了训练和检测的时间开销，从表3和表4中可以发现，本文提出的方法在训练时间上高于其他3种方法，特别是在CICIIDS2017数据中，参数为$\beta=2$，$\omega=0.26$，$\alpha_s=0.017$时，训练时间高达44 s，这是因为矩阵分解过程迭代次数较多，而遗传算法在参数优化的过程中需要较多的时间进行计算，但模型参数一旦优化后，利用生成的模型进行检测时，检测速度就能得到较大提高.

**表3 训练时间** /s

| 方法 | PCA | HMM | NMF | Alpha-NMF-GP | | |
|---|---|---|---|---|---|---|
| | | | | $\beta=2$, $\omega=0.45$, $\alpha_s=0.032$ | $\beta=0.5$, $\omega=0.3$, $\alpha_s=0.02$ | $\beta=2$, $\omega=0.26$, $\alpha_s=0.017$ |
| sendmail | 18 | 33 | 17 | 31 | 35 | 40 |
| CICIDS2017 | 23 | 37 | 21 | 38 | 42 | 44 |

**表4 检测时间** /s

| 方法 | PCA | HMM | NMF | Alpha-NMF-GP | | |
|---|---|---|---|---|---|---|
| | | | | $\beta=2$, $\omega=0.45$, $\alpha_s=0.032$ | $\beta=0.5$, $\omega=0.3$, $\alpha_s=0.02$ | $\beta=2$, $\omega=0.26$, $\alpha_s=0.017$ |
| sendmail | 16 | 14 | 12 | 9 | 11 | 8 |
| CICIDS2017 | 17 | 21 | 18 | 13 | 15 | 12 |

## 4 结　论

本文提出了一种应用遗传算法求解 Alpha-NMF 全局最优解的方法并用于入侵检测中. 为了确保矩阵分解过程的全局最优化，采用了 Alpha 散度来约束迭代更新规则，以实现最小化观测数据和模型之间的差异. 实验证明，Alpha-NMF 的迭代更新方式在入侵检测问题上是有效的，但它的学习过程较慢，因此，遗传算法可以用来寻找全局最优参数，这大大提高了检测方法的性能.

**参考文献：**


[1] 席荣荣，云晓春，张永铮，等. 一种改进的网络安全态势量化评估方法 [J]. 计算机学报，2015，38(4)：749-758.

[2] JAJODIA S，ALBANESE M. An Integrated Framework for Cyber Situation Awareness [J]. Theory and Models for Cyber Situation Awareness，2017，10030：29-46

[3] LOUKAS G，VUONG T，HEARTFIELD R，et al. Cloud-Based Cyber-Physical Intrusion Detection for Vehicles Using Deep Learning [J]. IEEE Access，2018，6(1)：3491-3508.

[4] AHMADIAN RAMAKI A，RASOOLZADEGAN A，JAVAN JAFARI A. A Systematic Review on Intrusion Detection Based on the Hidden Markov Model [J]. Statistical Analysis and Data Mining：the ASA Data Science Journal，2018，11(3)：111-134.

[5] 高　妮，高　岭，贺毅岳，等. 基于自编码网络特征降维的轻量级入侵检测模型 [J]. 电子学报，2017，45(3)：730-739.

[6] 徐慧敏，陈秀宏. 图正则化稀疏判别非负矩阵分解 [J]. 智能系统学报，2019，14(6)：1217-1224.

[7] MENG Y，SHANG R H，JIAO L C，et al. Feature Selection Based Dual-Graph Sparse Non-Negative Matrix Factorization for Local Discriminative Clustering [J]. Neurocomputing，2018，290：87-99.

[8] 钟　琴. 非负矩阵最大特征值的新界值 [J]. 西南大学学报(自然科学版)，2018，40(2)：40-43.

[9] LIAO Q，ZHANG Q. Local Coordinate Based Graph-Regularized NMF for Image Representation [J]. Signal Processing，2016，124：103-114.

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

[11] CICHOCKI A，ZDUNEK R，AMARI S I. Csiszár's Divergences for Non-Negative Matrix Factorization：Family of New Algorithms [M]//Independent Component Analysis and Blind Signal Separation. Berlin：Springer，2006：32-39.

[12] CICHOCKI A，CRUCES S，AMARI S I. Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization [J]. Entropy，2011，13(1)：134-170.

[13] DHILLON I S，SRA S. Generalized Nonnegative Matrix Approximations with Bregman Divergences [J]. Advances in Neural Information Processing Systems，2005(11)：5-8.

[14] 荣德生，段志田，胡举爽，等. 基于二阶锥规划与改进遗传算法的配网重构 [J]. 控制工程，2019，26(2)：223-228.

[15] 李建国，张海飞，周璐婕，等. 基于改进遗传算法的立体车库布局对比及服务资源优化 [J]. 西南大学学报(自然科学版)，2019，41(4)：139-148.

[16] BENJAMIN R，BARTLEY R. Sequence Aggregation Rules for Anomaly Detection in Computer Network Traffic [C]//American Statistical Association 2018 Symposium on Data Science and Statistics. Berlin：Springer，2018：1-13.

[17] VIJAYANAND R，DEVARAJ D，KANNAPIRAN B. Intrusion Detection System for Wireless Mesh Network Using Multiple Support Vector Machine Classifiers with Genetic-algorithm-based Feature Selection [J]. Computers & Security，2018，77：304-314.

[18] NICHOLAS L，OOI S Y，PANG Y H，et. al. Study of Long Short-Term Memory in Flow-Based Network Intrusion Detection System [J]. Journal of Intelligent & Fuzzy Systems，2018，35(3)：1-11.

[19] SHARAFALDIN I，HABIBI LASHKARI A，GHORBANI A A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization [C]//Proceedings of the 4th International Conference on Information Systems Se-

curity and Privacy. Funchal: Science and Technology Publications, 2018: 1-18.

[20] KUMAR N, MALLICK P. Blockchain Technology for Security Issues and Challenges in IoT [J]. Procedia Computer Science, 2018, 132: 1815-1823.

# An Intrusion Detection Model Based on Nonnegative Matrix Factorization Optimized by Genetic Programming

WU Yu[1], PENG Mao-ling[2],
QIAN Ren-fei[3], MA Zhe-feng[3], CHEN Shan-xiong[4]

1. *Chongqing Police College, Chongqing 401331, China;*
2. *Chongqing City Management College, Chongqing 401331, China;*
3. *Ningbo Dafa Chemical Fiber Company Limited, Ningbo Zhejiang 315336, China;*
4. *College of Computer and Information Science, Southwest University, Chongqing 400715, China*

**Abstract**: In the paper, nonnegative matrix factorization has been used to detect privileged program behavior of system monitoring, and then abnormal intrusion behavior been found. In the process of matrix factorization, Alpha-divergence is used as the measurement standard, and genetic programming algorithm is introduced to optimize the error function in non-negative matrix factorization. In the experiment, we have tested the system call data obtained from the Sun SPARC workstation of the University of New Mexico and CICIDS2017 data set. The experiment results show that our method has good performance in intrusion detection for detection accuracy.

**Key words**: intrusion detection; system call; nonnegative matrix factorization; genetic programming

责任编辑 张 枸