文章编号: 1001-0920(2019)02-0243-09 DOI: 10.13195/j.kzyjc.2018.1030

# 基于模糊粗糙集属性约简与GMM-LDA最优聚类簇特征学习的自适应网络入侵检测

刘金平[1,2], 张五霞[1], 唐朝晖[3], 何捷舟[1], 徐鹏飞[1†]

(1. 湖南师范大学 信息科学与工程学院，长沙 410081；2. 湖南师范大学 计算与随机数学教育部重点实验室，长沙 410081；3. 中南大学 信息科学与工程学院，长沙 410083)

**摘 要:** 网络入侵方式已日趋多样化,其隐蔽性强且变异性快,开发灵活度高、适应性强的实时网络安全监测系统面临严峻挑战. 对此, 提出一种基于模糊粗糙集属性约简 (FRS-AR) 和 GMM-LDA 最优聚类簇特征学习 (GMM-LDA-OCFL) 的自适应网络入侵检测 (ANID) 方法. 首先,引入一种基于模糊粗糙集 (FRS) 信息增益率的属性约简 (AR) 方法以实现网络连接数据最优属性集选择;然后,提出一种基于GMM-LDA的最优聚类簇特征学习方法,以获得正常模式特征库和入侵模式库的最优特征表示,同时引入模式库自适应更新机制,使入侵检测模型能够适应网络环境动态变化. KDD99 数据集和基于Nidsbench的网络虚拟仿真实验平台的入侵检测结果表明,所提出的ANID方法能有效适应网络环境动态变化,可实时检测出真实网络连接数据中的各种入侵行为,其性能优于当前常用的入侵检测方法,应用前景广阔.

**关键词:** 入侵检测；高斯混合模型聚类；模式匹配；模糊粗糙集；信息增益；模式更新

**中图分类号:** TP391.4 **文献标志码:** A

# Adaptive network intrusion detection based on fuzzy rough set-based attribute reduction and GMM-LDA-based optimal cluster feature learning

*LIU Jin-ping*[1,2], *ZHANG Wu-xia*[1], *TANG Zhao-hui*[3], *HE Jie-zhou*[1], *XU Peng-fei*[1†]

(1. College of Information Science and Engineering，Hu'nan Normal University，Changsha 410081，China；2. Key Laboratory of Computing and Stochastic Mathematics，Ministry of Education，Hu'nan Normal University，Changsha 410081，China；3. School of Information Science and Engineering，Central South University，Changsha 410083，China)

**Abstract:** With the increasing diversity and rapid variability of network intrusion, the development of real-time network security monitoring systems with high flexibility and strong adaptability still faces severe challenges. Therefore adaptive network intrusion detection (ANID) method based on fuzzy rough set attribute reduction (FRS-AR) and Gaussian mixture model linear discriminant aualysis (GMM-LDA) optimal cluster feature learning (GMM-LDA-OCFL) is proposed. Based on the fuzzy rough set theory, the optimal attribute set of network connection data is selected automatically by information gain rate measurement. Then, an optimal cluster feature learning method based on GMM-LDA is proposed to obtain the optimal feature representation of the normal mode feature library and the intrusion mode feature library. At the same time, the adaptive on-line update mechanism of the normal (abnormal) pattern feature library is introduced, so that the detection model can adapt itself to dynamic network changes. The test results of KDD99 and network simulation experiment platform based on Nidsbench show that the proposed method can effectively adapt to the dynamic changes of the network environment and various intrusion behaviors in the real network connection data can be detected in real time. And the performance of the proposed method is better than that of the existing commonly-used intrusion detection methods, which has potentially wide application prospects.

**Keywords:** intrusion detection; GMM clustering; pattern matching; fuzzy rough set; information gain; model updating

**收稿日期:** 2018-07-26；**修回日期:** 2018-09-18.

**基金项目:** 国家自然科学基金项目(61501183, U1701261, 61771492)；湖南省自然科学基金项目(2018JJ3349)；图像信息处理与智能控制教育部重点实验室(华中科技大学)开放基金项目(IPIC2017-03).

**责任编委:** 阳春华.

†通讯作者. E-mail: xupf@hunnu.edu.cn.

## 0 引 言

随着经济全球化和信息全球化的迅猛发展,人们在享受互联网的普及所带来便捷的同时,也面临着日益严重的网络安全威胁. 网络入侵检测[1]作为一种主动的安全防护技术,已经成为安全防护中必不可少的一部分. 常用的入侵检测技术可以分为误用检测[2]和异常检测[3]. 由于网络入侵者(黑客)不断升级其方法和技术,目前传统的入侵检测方法,无论是误用检测还是异常检测,都难以获得令人满意的结果. 因此,如何迅速有效地发现各类新的入侵行为,对于保证计算机网络安全访问变得极为重要.

在网络入侵检测中,由于正常(异常)网络连接内部具有较强的相似性,而入侵行为与正常连接数据之间存在较大的差异性,可以将异常检测与误用检测相结合,采用基于聚类的模式挖掘方法,分别对正常和异常连接数据进行类别划分[4]以获得辨识度高的特征模式库,再通过模式匹配等方法实现入侵报警. 本文基于该思路进行网络入侵检测,其中一个关键性环节为正常(异常)网络连接的聚类簇特征库学习.

常用的聚类方法[5]包含以下几类:基于划分的聚类、基于层次的聚类、基于密度的聚类、基于网格的聚类和基于模型的聚类等.

$K$-means[6]是一种经典的划分聚类方法. 该算法以误差平方和作为度量聚类质量的目标函数,易于理解和实现. 但该算法需人工确定聚类个数,且聚类结果对于初始聚类中心的选取有严重的依赖性.

层次聚类[7]中距离和规则的相似度容易定义,可以较好发现类的层次关系. 但算法复杂,对数据奇异值敏感度高,且聚类层次的选择依赖于实际数据客观结构及人工经验,造成聚类结果不稳定[8].

基于密度的DBSCAN算法[9]理论上可以发现任意形状的聚类簇,但对引入的密度阈值及用于快速收敛到高密度区域的自适应密度值非常敏感.

基于网格的方法[10]执行速度只依赖于数据空间中每个维度上数据单元的个数,执行效率高. 但该方法对参数敏感,难以处理不规则分布数据,且对高维数据难以获得较好的结果.

基于模型的方法[11]为每个聚类假定一个模型,寻找能够很好满足该模型的数据集. 这类方法往往不是将数据严格归属于某一类,而以概率形式表现. 每一类的特征也可用参数来表达,可以达到较好的聚类效果. 但该方法同样需要事先给定聚类数目.

综上所述,当前绝大多数的聚类算法均需凭经验手动确定聚类数目,且最终聚类结果严重依赖于初始聚类中心的选择. 为获得辨识度高、稳定性强的网络入侵检测特征库,避免聚类算法对初始聚类中心过分依赖、聚类个数较难确定等问题,本文提出一种基于高斯混合模型(GMM)与线性辨别分析(LDA)相结合(GMM-LDA)的最优聚类簇特征学习(GMM-LDA-OCFL)方法. 该方法能自动确定聚类个数,避免因聚类个数、初始聚类中心及相关参数设置不准确,给网络连接特征库学习带来不利影响.

以上算法主要是针对低维特征空间的,在高维条件下易产生“维度灾难”. 在网络入侵检测中,网络连接数据的特征维数高,数据特征具有较大的冗余性和不确定性,会极大影响入侵检测结果的有效性. 因此,本文将模糊粗糙集(FRS)理论引入到网络入侵检测的属性约简(AR)中. 在约简属性空间对正常模式和异常模式特征库进行聚类学习,最终提出一种基于FRS-AR和GMM-LDA-OCFL的自适应网络入侵检测(ANID)方法.

本文提出的ANID方法主要创新点表现在: 1)基于FRS理论,以信息增益率为准则实现网络连接数据本征特征自动选取,有效提升系统整体效率和入侵检测性能的稳定性; 2)基于GMM-LDA最优聚类簇特征学习,获得正常模式特征库和入侵模式的最优特征表示,该方法无需人工确定聚类个数,可有效避免聚类个数及聚类中心初始值对特征库学习的影响; 3)引入正常(异常)模式特征库自适应在线更新策略,使入侵检测模型能够自适应动态网络变化,对已知和未知入侵都具备较高的检测率.

## 1 自适应网络入侵检测

本文的ANID方法主要包含四大模块: (A)属性约简与模式表示; (B)模式库生成; (C)模式匹配与更新; (D)报警与响应. 具体流程如图1所示.

### 1.1 属性约简与模式表示

由于原始数据往往包含一些隐含信息,本文利用信息函数将这些隐含信息提取出来,在保留原始特征的同时更好地表现数据特征[12]. 将网络连接记录表示为四元组$\mathrm{FS} = \langle U, A_t, V, f\rangle$. 其中:$U$表示整个网络数据集; $A_t$是一个非空的有限属性集;属性域集合$V = U_{a\in A_t}V_a$, $V_a$是属性$a$的值集合,称为$a$的域; $f: U\times A_t \to V$是信息函数,它是从属性域向对象分配特定的值.

属性集合$A_t$中,设置条件属性集合$C$(除标签属性外的所有属性的集合), $D$是决策属性集合,于是上述四元组可表示为

$$\mathrm{FDS} = \langle U, C\bigcup D, V, f\rangle. \tag{1}$$

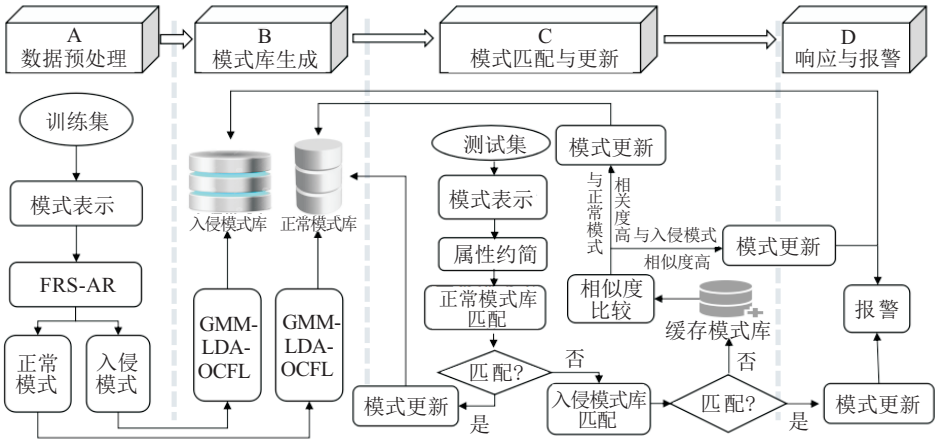


**图 1　基于FRS-AR与GMM-LDA-OCFL的ANID方法流程**

### 1.1.1　FRS理论

基于粗糙集理论可以获得保留原始特征辨别能力的属性子集[13]. 然而,经典粗糙集理论只能处理离散属性集,不能很好地处理包含大量连续值的网络连接数据.

本文引入FRS理论[12],基于FRS信息增益率对网络连接数据特征属性集进行自动选取,以获得约简的本征属性集,从而最终提高入侵检测算法的稳定性和有效性. 下面从FRS理论的基本定义[12]出发,引出基于FRS增益率测量的AR算法.

模糊等价关系是FRS的核心. 给定非空有限集合$X$,$X$上的模糊等价关系$R$可用关系矩阵$M(R)$表示,即

$$H_x=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}. \tag{2}$$

其中:$r_{ij}\in[0,1]$是$x_i$与$x_j$的关系值;$x_i$和$x_j$分别表示不同数据在同一属性上的值,$x_i,x_j\in X$. 模糊等价关系需满足自反、对称和传递性. 本文采用如下函数计算$x_i$与$x_j$的关系值:

$$r_{ij}=\begin{cases} 1-4\times\dfrac{|x_i-x_j|}{|a_{\max}-a_{\min}|}, & \dfrac{|x_i-x_j|}{|a_{\max}-a_{\min}|}\leqslant 0.25; \\ 0, \text{ otherwise.} \end{cases} \tag{3}$$

其中:$a_{\max}$和$a_{\min}$分别表示属性$a$的最大值和最小值.

**定义1**　数据集$U$根据$R$模糊划分可定义为

$$U/R=\{[x_i]_R\}_{i=1}^n;$$
$$\text{s.t. } [x_i]_R=\frac{r_{i1}}{x_1}+\frac{r_{i2}}{x_2}+\cdots+\frac{r_{in}}{x_n}. \tag{4}$$

其中:$U/R$是模糊划分,$[x_i]_R$是模糊集合.

**定义2**　模糊等价关系的信息量定义为

$$H(R)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_R|}{n}, \tag{5}$$

其中$|[x_i]_R|=\sum\limits_{j=1}^{n}r_{ij}$表示集合基数.

**定义3**　给定一个模糊信息系统FIS $=\langle U, C\bigcup D,V,f\rangle$. $P$和$Q$是属性$C$的两个子集,$[x_i]_P$和$[x_i]_Q$分别表示被$P$和$Q$划分的包含$x_i$的模糊等价类. $P$和$Q$的联合熵定义为

$$H(PQ)=H(R_PR_Q)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_P\bigcap[x_i]_Q|}{n}. \tag{6}$$

**定义4**　在模糊决策系统FIS $=\langle U,C\bigcup D,V,f\rangle$中,设$B$为$C$的子集,则$D$在$B$条件下的条件熵定义为

$$H(D\mid B)=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_B\bigcap[x_i]_D|}{|[x_i]_B|}. \tag{7}$$

**定义5**　给定一个模糊信息系统FIS $=\langle U,C\bigcup D,V,f\rangle$,$B$为$C$的子集,$B$和$D$的互信息定义为

$$I(B;D)=H(D)-H(D\mid B). \tag{8}$$

### 1.1.2　FRS-AR

给定一个模糊决策系统FIS $=\langle U,C\bigcup D,V,f\rangle$,设$B\subseteq C,\forall a\in C-B$,属性$a$的信息增益Gain$(a,B,D)$[12]可以表示为

$$\text{Gain}(a,B,D)=I(B\bigcup\{a\};D)-I(B;D); \tag{9}$$

属性$a$的信息增益率$\text{Gain}_{\text{Ratio}}(a,B,D)$定义为

$$\text{Gain}_{\text{Ratio}}(a,B,D)=\frac{\text{Gain}(a,B,D)}{H(a)}=\frac{I(B\bigcup\{a\};D)-I(B;D)}{H(a)}, \tag{10}$$

$\text{Gain}_{\text{Ratio}}(a,B,D)$可用来衡量属性$a$的重要程度.

基于FRS属性约简,可通过每次选择FRS增益率最大的特征进行属性集选取,最终获得的属性集即为

约简的本征属性集. 主要步骤如下所示.

**算法1** FRS-AR.

输入: 数据集$X$、条件属性集$C$、决策属性集$D$;

输出: 约简的属性集$B$.

Step 1: 置空$B$集合;

Step 2: 对于每个属性$a \in C-B$,计算信息增益率$\mathrm{Gain}_{\mathrm{Ratio}}(a,B,D)$;

Step 3: 筛选$\mathrm{Gain}_{\mathrm{Ratio}}(a,B,D)$的最大值,并将对应的属性定义为$b$;

Step 4: 如果$\mathrm{Gain}_{\mathrm{Ratio}}(a,B,D)$的最大值大于0,则$B \leftarrow B \bigcup \{b\}$,返回Step 2,否则继续Step 5;

Step 5: 集合$B$即为属性约简后的属性集合.

### 1.2 基于GMM-LDA-OCL的离线模式库生成

GMM聚类[14]方法是一种基于概率密度模型的聚类方法. 设$x \in \boldsymbol{R}^d$,基于GMM,$x$的概率密度$p(x)$可以采用$k$个高斯随机变量的加权模型来表示,每个高斯模型代表一个聚类. GMM模型[14]可表示为

$$p(x)=\sum_{k=1}^{K}\pi_k N(x \mid \mu_k,\Sigma_k). \tag{11}$$

其中: $K$为高斯模型个数,$\pi_k$为第$k$个高斯模型的权重,$\mu_k$和$\Sigma_k$分别为第$k$个高斯模型的均值与方差.

GMM参数可以采用最大期望算法(EM)进行迭代求解. 将GMM模型参数重写为$\Omega = \{\pi_k,\mu_k,\Sigma_k\}_{k=1}^{K}$,GMM模型参数更新规则如下:

$$p(k \mid x_n,\Omega^{\mathrm{old}})=\frac{\pi_k^{\mathrm{old}}N(x_n \mid \mu_k^{\mathrm{old}},\Sigma_k^{\mathrm{old}})}{\sum\limits_{j=1}^{k}\pi_j^{\mathrm{old}}N(x_n \mid \mu_j^{\mathrm{old}},\Sigma_j^{\mathrm{old}})}. \tag{12}$$

其中: $1 \leqslant l \leqslant K, 1 \leqslant n \leqslant N$, $N$为样本个数. 于是,计算新的GMM模型参数$\Omega^{\mathrm{new}}=\{\pi_k^{\mathrm{new}},\mu_k^{\mathrm{new}},\Sigma_k^{\mathrm{new}}\}_{k=1}^{K}$的更新规则如下:

$$\mu_k^{\mathrm{new}}=\sum_{k=1}^{K}x_n p(k \mid x_n,\Omega^{\mathrm{old}})\Big/\sum_{k=1}^{K}p(k \mid x_n,\Omega^{\mathrm{old}}), \tag{13}$$

$$\Sigma_k^{\mathrm{new}}=\frac{1}{N}\sum_{n=1}^{N}p(k \mid x_n,\Omega^{\mathrm{old}})(x_n-\mu_k^{\mathrm{new}})^{\mathrm{T}}(x_n-\mu_k^{\mathrm{new}}), \tag{14}$$

$$\pi_k^{\mathrm{new}}=\frac{1}{N}\sum_{k=1}^{K}p(k \mid x_n,\Omega^{\mathrm{old}}). \tag{15}$$

#### 1.2.1 LDA

GMM聚类需要事先给定聚类个数$K$,且研究表明,在聚类过程中,模型参数的求解受各高斯模型的初值影响,不恰当的模型初值极易造成式(13)~(15)的迭代过程难以收敛,从而难以保证有效的聚类结果[15].

本文采用基于LDA[16]的信息增益算法来自动确定聚类个数$K$. LDA充分考虑了分类标签信息,寻求投影后不同类别之间数据点距离最大化以及同一类别数据点间距离最小化. 因此,利用LDA可以获得最好的聚类性能.

假设数据集$D=\{(x_1,y_1,),\cdots,(x_n,y_n,)\}$,$x_i$是经数据预处理后的$d$维向量,$y_i \in \{C_1,C_2,\cdots,C_k\}$. 经GMM聚类后,设$\mu_j(j=1,2,\cdots,K)$为第$j$类样本的均值向量,$\Sigma_j(j=1,2,\cdots,K)$为第$j$类样本的协方差矩阵. 设投影空间是一个$m$维的超平面,对应的基向量为$(\omega_1,\omega_2,\cdots,\omega_m)$,基向量组成的矩阵为$W \in \boldsymbol{R}^{d\times m}$.

根据LDA算法求投影后不同类别之间数据点距离更大化以及同一类别数据点距离最小化原理,本算法的优化目标为

$$J(W)=\frac{\sum\limits_{k=0}^{K-1}W^{\mathrm{T}}\Sigma_k W}{\sum\limits_{i=0}^{K-1}\sum\limits_{j=0}^{K-1}W^{\mathrm{T}}(\mu_i-\mu_j)(\mu_i-\mu_j)^{\mathrm{T}}W}. \tag{16}$$

由于类内散度矩阵$S_W$可表示为

$$S_W=\sum_{j=1}^{K}S(\omega_j)=\sum_{j=1}^{K}\sum_{x\in x_j}(x-\mu_j)(x-\mu_j)^{\mathrm{T}}, \tag{17}$$

类间散度矩阵$S_b$为

$$S_b=\sum_{j=1}^{K}N_j(\mu_j-\mu)(\mu_j-\mu)^{\mathrm{T}}. \tag{18}$$

其中$\mu$为所有样本均值向量. 因此,优化目标$J(W)$可重写为

$$J(W)=\frac{W^{\mathrm{T}}S_W W}{W^{\mathrm{T}}S_b W}. \tag{19}$$

由于$W^{\mathrm{T}}S_b W$和$W^{\mathrm{T}}S_W W$都是矩阵,不是标量,无法作为一个标量函数来优化,因此,这里将目标函数改写为

$$J(W)=\frac{W^{\mathrm{T}}S_W W}{W^{\mathrm{T}}S_b W}=\frac{\prod\limits_{i=1}^{d}\omega_i^{\mathrm{T}}S_W\omega_i}{\prod\limits_{i=1}^{d}\omega_i^{\mathrm{T}}S_b\omega_i}=\frac{\prod\limits_{i=1}^{d}\omega_i^{\mathrm{T}}\omega_i}{\prod\limits_{i=1}^{d}\omega_i^{\mathrm{T}}S_bS_W^{-1}\omega_i}. \tag{20}$$

此时,$J(W)$的最小值为矩阵$S_bS_W^{-1}$的最大特征值,最大值是矩阵$S_bS_W^{-1}$的最小特征值.

#### 1.2.2 GMM-LDA

GMM-LDA聚类先采用GMM对已知数据库的正常数据和入侵数据分别进行聚类,然后对每次

GMM聚类结果采用LDA 算法进行测量. 根据目标函数$J(W)$对当前聚类个数$K$效果进行评估. 依次迭代,直至目标函数收敛或达到最小值时,输出聚类结果,结束算法.

聚类个数$K$由式(20)中的目标函数$J(W)$来确定. 对于每次迭代的聚类个数$K$,当$J(W)$随着$K$值变化达到最小值时,即为最佳聚类个数. 同时,结合更新模式逐渐动态生成最近的正常模式库以及入侵模式库. 更新模式将在2.3 节描述.

基于GMM-LDA聚类的最优聚类簇特征学习的主要步骤如下.

**算法2**　基于GMM-LDA最优聚类簇特征学习.

Step 1:输入数据集$D$和初始聚类个数$K$,有

$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\},$$

$$y_i\in\{C_1,C_2,\cdots,C_k\}.$$

Step 2: 获得$D$的GMM聚类模型;通过式(13)～(15)进行迭代更新获得GMM聚类模型参数$\{\pi_k,\mu_k,\Sigma_k\}_{k=1}^{K}$.

Step 3: 采用式(20)所示的LDA模型计算目标函数最小值$J(W)$.

Step 4: 将当前的目标函数最小值$J(W)$与上一次的值进行比较,如果小于上次值,则继续增加聚类个数,并返回Step 2;否则,说明已经得到最小目标函数值(目标函数值收敛),执行Step 5.

Step 5: 停止聚类,并保留此时聚类个数$K$,此时得到的聚类个数$K$即为最终的最优聚类个数.

### 1.3　模式匹配与更新以及响应与报警

针对网络环境动态变化的特性,为了提高网络入侵检测模型在动态环境下的自适应性,本文引入一种模式自动更新机制.

通过新建模式缓存库,将与正常模式库和入侵模式库中都不匹配的数据加入到该模式缓存库. 模式缓存库中每一个模式都有一个生命值,当一条新记录匹配这个模式时,该模式的生命值加1,同时其他模式的生命值将衰减$\varepsilon(\varepsilon\ll 1)$. 当该模式生命值达到阈值$\tau$时,通过分析该模型与正常模式或异常模式的相似度来确定将该模式加入正常模式库或入侵模式库进行模式动态更新. 模式匹配与更新以及响应与报警流程如图2所示.

在模式匹配与更新中,采用余弦相似度来测量待检测对象与相应模式的相异度以进行模式匹配. 余弦相似度计算方法为

$$d(x,y)=\frac{x^{\mathrm{T}}y}{\|x\|\|y\|},\tag{21}$$

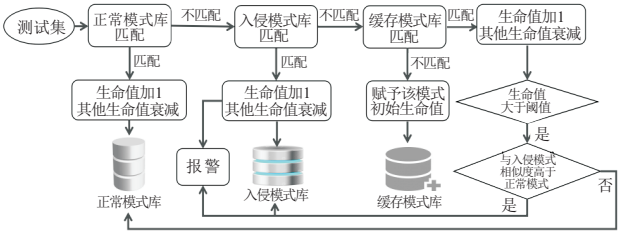


**图 2　模式匹配与更新以及响应与报警**

其中$d(x,y)$表示待检测量样本$x$与模式$y$之间的余弦匹配度,该值越大表明匹配度越高(代表$x$与$y$间的夹角越小).

### 1.4　基于FRS-AR与GMM-LDA-OCFL的ANID

本文所提出的ANID算法的主要流程如下.

初始化:给定训练集$D$、测试集$X$、模式衰减系数$\varepsilon$、频繁模式阈值$\tau$;正常模式库、入侵模式库、模式缓存库初始都为空.

Step 1: 采用算法1进行最优属性集提取.

Step 2: 采用算法2分别进行正常模式库和入侵模式库学习.

Step 3: 在正常模式库中搜索与测试集数据$X$相匹配的模式. 在搜索的同时累加所有正常模式对$X$的相似度值得到$F_n$,并对模式的生命值进行衰减. 若在正常模式库中发现与$X$匹配的模式,则对该模式进行模式更新,生命值加1,$X$标记为正常模式,重新执行Step 3对新样本进行检测;若不匹配,则执行Step 4.

Step 4: 在入侵模式库中搜索与$X$匹配的模式,在搜索的同时累加所有入侵模式对$X$的相似度值,得到$F_i$,并对模式的生命值进行衰减. 若发现与$X$匹配的模式,则报警,同时进行模式更新,生命值加1,结束对此记录的处理,并回到Step 3对新样本进行检测;若不匹配,则进行Step 5.

Step 5: 在模式缓存库中搜索与$X$匹配的模式,并对所有模式的生命值衰减. 若发现与$X$匹配的模式,则进行模式更新,生命值加1,并作出正常/报警响应;若不匹配,则比较相似度$F_n$和$F_i$,当$F_n>F_i$时,标记为正常模式,并对该模式进行更新;否则,标记为入侵模式,报警并进行模式更新.

Step 6: 当缓存库中模式的生命值达到阈值$\tau$时,将该模式加入相应的(正常/入侵)模式库,结束对此记录的处理,返回Step 3继续对新样本进行检测.

## 2　实验验证

为验证本文所提出的ANID方法的性能,分别在KDD99数据集上和基于Nidsbench的网络仿真平台

上进行实验.

### 2.1 KDD99数据集实验

KDD99是由MIT Lincoln实验室提供的1998 DARPA入侵检测评估数据集的一个扩充版本,包括训练集和测试集. 数据以网络连接的形式保存,完整的训练数据集包含4 999 000条链接记录,每条记录由42个属性组成,其中最后一个属性为该记录的类型标签. 这些入侵行为被分成4大类: U2R, R2L, DOS和PROBING,整个训练集中的入侵类型及数目如表1所示.

**表1 特征提取方法比较**

| 特征提取方法 | 提取属性个数 | 检测率 | 误报率 |
|---|---|---|---|
| 本文方法 | 20 | 0.921 0 | 0.001 6 |
| KPLS[17] | 23 | 0.872 1 | 0.003 1 |
| PCA-ICA[18] | 19 | 0.890 3 | 0.002 0 |
| Rank-KPCA[19] | 32 | 0.790 1 | 0.002 3 |
| PLS[20] | 22 | 0.788 2 | 0.003 7 |

本文以10%的KDD 99数据子集作为实验数据,训练集中共包含494 019条记录,其中,正常行为记录97 276条,入侵行为记录396 743条,入侵行为共22种;测试数据使用带标识的测试数据集,该数据集共311 029条记录,其中,正常行为记录60 593条,入侵行为记录250 436条,入侵行为共37种,这些入侵行为中包含训练数据中的20种入侵行为和17种训练数据中未出现的入侵行为.

在测试样本集时,共选取4组数据,每组各3万条记录,为确保实验结果的准确性,每组正常数据和入侵数据分别为28 000和2 000. 其中: 第1组和第2组测试集选自训练集,第3组和第4组选自KDD 99数据中除去训练集的部分. 为了更好地测试对未知入侵的检测能力,在选取时特别选取了一些没有包含在训练集中的入侵,即未知的入侵. 同时为了验证系统的可伸缩性,使用训练集的全集进行实验比较.

#### 2.1.1 数据预处理

1) 数据标准化.

为了提高算法精度和加快算法的收敛速度,通常需要将数据标准化.

i) 计算平均的绝对偏差$S_f$,即

$$S_f = \frac{1}{N}\sum_{i=1}^{N}|X_{if} - m_f|. \tag{22}$$

其中: $x_{1f}, x_{2f}, \cdots, x_{nf}$为$f$的$N$个属性特征值; $m_f$是$f$的均值,即

$$m_f = \frac{1}{N}\sum_{i=1}^{N}X_{if}. \tag{23}$$

ii) 计算标准化后的特征属性值

$$Z_{if} = \frac{x_{if} - m_f}{S_f}. \tag{24}$$

2) 属性约简.

采用FRS-AR方法进行最优特征提取. 根据特征间的关系以及特征与决策属性之间的关系,选出合适的特征子集. 将本文方法与其他特征提取方法进行比较,结果如表1所示.

表1的结果表明: 虽然PCA-ICA可以将数据维数降到19维,但其检测率降低,误报率较高;而本文所采用的属性约简方法,通过选择FRS增益率高的20个属性,在降低维度的同时,仍然可以获得较高的检测率和较低的误报率. 由此可见,FRS-AR方法的综合性能高于其他算法.

#### 2.1.2 频繁模式阈值对ANID性能的影响

在模式更新时,需要给定频繁模式阈值$\tau$以确定某个新的样本是否属于频繁模式,进而进行模式库更新. 在测试时,为了尽量保持学习时得到的知识,取模式衰减参数$\varepsilon = 0.000\,1$,同时选取不同的频繁模式阈值$\tau$(取值分别为50、100、1 000),并将实验自始至终分成8个阶段,分别测试频繁模式阈值$\tau$的不同取值对实验不同阶段的影响. 实验结果如图3所示.

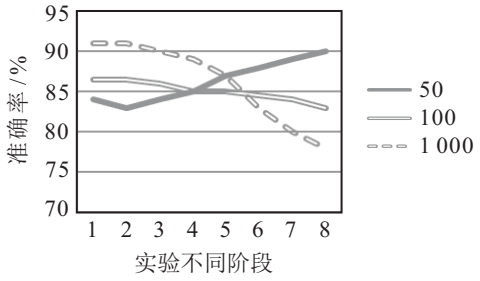


**图3 不同$\tau$值实验结果**

图3的结果表明,在实验的不同阶段,阈值$\tau$对实验的准确率和误报率的影响不同. 在初始阶段,较高的频繁阈值准确率更高;随着实验的进行,频繁阈值越低,准确率越高. 因此,为了获得较高的检测率和较低的误报率,在实验初始阶段,由于各模式库中模式较少,待测试数据量较大,可以先将$\tau$设置得较大;随着实验的进行,正常模式库和入侵模式库中的模式逐渐增多,大多数数据与正常模式库或入侵模式库的匹配概率明显增大,与模式缓存库匹配变成了小概率事件. 此时阈值$\tau$可以根据流入模式缓存库的数据量的减少而自动减小.

#### 2.1.3 实验结果与分析

1) 验证性结果.

采用两个评估指标进行评估,分别是检测率和误检率,检测结果如表2所示. 由表2中可以看出: 平均检测率达到93.3%,最高达到97.4%;同时,误检率保

持较低水平,平均值在0.25 %以下. 从而表明该网络入侵检测方法具有较好的性能.

**表 2　实验结果**

| 数据集 | 正常实例数 | 入侵实例数 | 检测率/% | 误报率/% |
|---|---|---|---|---|
| 1 | 25 650 | 1 410 | 90.2 | 0.12 |
| 2 | 27 800 | 1 420 | 97.4 | 0.20 |
| 3 | 26 310 | 1 080 | 91.3 | 0.24 |
| 4 | 26 850 | 1 380 | 94.1 | 0.17 |
| 平均值 | 26 653 | 1 332 | 93.3 | 0.18 |

表3显示了本文所提出的自适应网络入侵检测方法在4组测试集上对已知和未知攻击入侵的检测情况. 由表3的检测结果可以看出,本文提出的方法在4组测试集中对R2L入侵类型的检测准确率较低. 这是因为该入侵类型与某些正常类型的相似度较高,可以伪装成正常模式进行攻击,导致检测困难. 但本文所提出的方法对未知入侵的检测率都超过了89 %,对未知入侵的检测准确率最高可以达到97.7 %. 综合所有数据的检测结果,本文所提出的ANID方法不仅对已知入侵类型能够达到高检测率和低误报率,对未知入侵也能达到很好的检测效果.

**表3　4组数据集对已知和未知入侵的检测结果**　%

| 类别 | 第1组 | | 第2组 | | 第3组 | | 第4组 | |
|---|---|---|---|---|---|---|---|---|
| | 已知入侵 | 未知入侵 | 已知入侵 | 未知入侵 | 已知入侵 | 未知入侵 | 已知入侵 | 未知入侵 |
| DOS | 86.2 | 90.1 | 98.5 | 93.2 | 88.5 | 87.6 | 94.8 | 91.2 |
| U2R | 92.2 | 93.1 | 99.2 | 97.2 | 95.3 | 88.4 | 93.3 | 92.6 |
| R2L | 96.4 | 89.9 | 98.2 | 96.3 | 96.5 | 90.3 | 96.4 | 92.1 |
| PROBING | 91.6 | 90.1 | 98.9 | 97.7 | 94.1 | 89.7 | 91.1 | 94.9 |
| 合计 | 91.6 | 90.8 | 98.7 | 96.1 | 93.6 | 89.0 | 93.9 | 92.7 |

2) 性能比较.

为了更好地测试基于聚类簇特征学习的网络入侵检测系统模型的性能,将本文所提出的ANID方法与一些经典的入侵检测方法进行比较. 这些对比方法的具体细节可参见对应的参考文献. 对比实验结果如表4所示.

**表4　在KDD 99数据集上的性能比较**

| 模型 | 检测率 | | 误报率 | |
|---|---|---|---|---|
| | 均值 | 方差 | 均值 | 方差 |
| 本文方法 | 0.932 5 | 0.077 6 | 0.001 8 | 0.007 7 |
| CANN[21] | 0.868 8 | 0.101 9 | 0.001 6 | 0.027 7 |
| HFSTE[22] | 0.891 8 | 0.091 6 | 0.007 6 | 0.012 3 |
| FBSLAIDS[23] | 0.888 3 | 0.131 8 | 0.013 1 | 0.014 6 |

表4是几种方法在4组数据集上的平均实验结果. 从误报率来看,CANN[21]比本文方法的误报率低(只低了0.02 %),但准确率较低. 综合检测率和误报率,本文的ANID方法性能更优. 其原因主要是因为CANN采用聚类中心与最近邻居方法相结合的新特征表示方法,利用两个距离度量作为最终属性进行入侵检测,虽然大大提高了系统效率,但同时也丢掉了数据的一些重要属性,从而影响最终的准确率. 而本文提出的ANID方法,综合了数据的各个属性及分类关系,保留了网络连接数据中最重要的属性,因而能获得更好的检测性能.

与HFSTE[22]比较,本文方法的检测率较高,同时误报率较低,因此,本文方法的入侵检测性能明显优于SDCMIDS方法. 本文方法检测率高是因为采用模式更新机制进行正常(入侵)模式库在线更新,能够自适应网络环境的发展变化.

与FBSLAIDS[23]进行比较,本文方法检测率较高,误报率较低. 需要指出的是,FBSLAIDS有个训练过程,而本文入侵检测系统在进行正常(入侵)模式库进行聚类学习之后,还可以对所获得的特征库进行实时更新,实时或近实时地运行.

总之,这些对比方法的入侵检测率均比本文方法要低超过4个百分点,本文方法具有明显的性能优势,可以获得更高的检测率和更低的误报率,且能够在线更新模型以适应网络环境的动态变化,具有广泛的应用前景.

### 2.2　Nidsbench仿真平台实验

为了更好地测试本文提出的ANID方法性能,本文采用Nidsbench 测试平台(Performance analysis of Inidsbench)[24]搭建一个网络仿真实验平台以进行网络入侵检测. 先对实际网络流量、网络主机使用和各种各样的攻击行为进行仿真,再根据仿真事件表和入侵检测结果生成评估报告. 仿真实验同样采用网络入侵检测率和误报率作为算法模型的评价标准.

Nidsbench是Anzen公司开发的IDS测试软件包,包括Tcpreplay和Fraqrouter两部分[24]. Tcpreplay为了还原当时网络实际运行状态,将系统截获的数据包

重放;Fraqrouter为了测试入侵检测的正确性和安全性,构造了一系列不易被入侵检测系统检测到的攻击. 通过Nidsbench测试平台可以实现模拟入侵,以测试自适应网络入侵检测方法的有效性.

仿真实验的拓扑结构如图4所示. 该仿真网络由3个路由器(router1、router 2、router 3,路由器间采用RIP协议),3个交换机(switch1、switch 2、switch 3),4个服务器(数据库服务器、Email服务器、Ftp服务器、Web服务器),6个以太子网(subnet 1,···,subnet 6)构成.

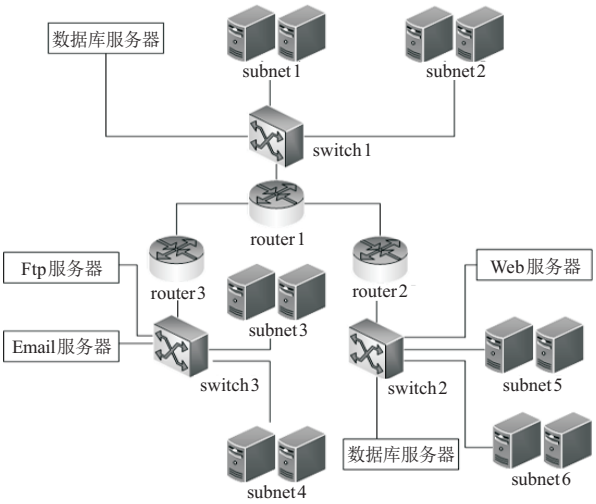


**图 4 仿真网络拓扑图**

### 2.2.1 数据收集

仿真实验是在一个内部网中仿真用户操作,模拟入侵,重建正常通信和攻击数据,用于自适应网络入侵检测方法的测试. 采用分时间段对实际流量进行分析的方法,通过统计计算得到各个协议流量概率分布,以此为模型,分别仿真各个协议的流量. 在网络中各个子网的每个主机都按照一定流量模型运行数据库服务、Email 服务、Ftp服务和http服务.

测试数据中包含训练数据中没有出现过的数据类型. 其中,正常数据和入侵数据均为42维数据(包含41维条件属性和1维决策属性),经过属性约简后可得到19维的数据特征. 训练数据帮助自适应网络入侵检测方法建立正常模式库和入侵模式库,调整各参数的设置. 实际测试数据中包含测试数据中没有的数据类型,用于测试检测方法.

### 2.2.2 仿真过程

首先,将各种正常网络连接和入侵活动的数据记录作为训练数据,对获得的训练数据归一化,采用本文所提出的基于FRS-AR 方法进行属性约简;然后,使用GMM-LDA算法对约简后的网络连接数据进行最优聚类簇特征聚类,生成正常模式库和入侵模式库;最后,在测试平台上重放用户的活动,为了测试对未知活动的检测效果,同时模拟训练数据中未出现的用户正常活动和入侵活动,根据准确率和误报率两个评估指标测试本文提出的ANID方法的有效性.

### 2.2.3 仿真结果

仿真实验结果如表5所示. 针对不同攻击行为分别进行检测,由检测结果可以看出,对U2L类型数据的检测率低于其他数据类型,这是因为本文方法是利用相似度判定数据类型,而该类型数据与正常模式数据的相似度较高,导致检测率和误报率不理想. 但从整体来看,仿真环境下的检测率超过90%,同时误报率仅为0.33%,表明本文方法是可行的,并且综合性能良好.

**表5 检测结果**

| 类别 | 样本数 | 类型 | 明细 | 准确率(TP)/% | 平均准确率/% | 误报率(FP)/% | 平均误报/% |
|---|---|---|---|---|---|---|---|
| 正常 | 14 700 | Normal | 10 523 | 92.7 | | 0.18 | |
| | | Probe | 1 714 | 91.4 | | 0.32 | |
| | | R2L | 2 463 | 91.3 | 90.1 | 0.13 | 0.33 |
| 攻击 | 5 300 | Vulnerability | 1 972 | 89.9 | | 0.24 | |
| | | Dos | 1 947 | 91.2 | | 0.52 | |
| | | U2R | 1 381 | 84.3 | | 0.61 | |

## 3 结 论

本文结合FRS-AR与GMM-LDA-OCL,提出了一种ANID方法. 该方法基于FRS理论,通过增益率测量对网络连接数据进行属性约简,采用GMM-LDA进行最优聚类簇特征学习. 该方法可以自动确定最优聚类数目,实现正常模式、入侵模式最优特征库的自动提取,有效改善了随机选取聚类个数导致聚类结果不稳定的缺陷,从而提高了系统的稳定性.

同时,本文引入了自适应模式更新机制,通过引入频繁模式发掘,实时更新正常模式和入侵模式聚类库,保证系统的自适应性,实现对各种攻击行为的检测,从而提高了网络入侵检测系统的性能.

在KDD99数据集上和Nidsbench网络虚拟仿真平台上进行了大量的验证性和对比性实验,结果表明,本文提出的ANID方法性能明显优于当前常用的入侵检测方法. 由于本文所提出的ANID方法能自适

应网络环境的变化,无论是针对已知的入侵类型还是未知入侵类型,均具有较高的检测率和较低的误报率,应用前景广阔.

## 参考文献(References)


[1] Baig M M, Awais M M, El-Alfy E S M. A multiclass cascade of artificial neural network for network intrusion detection[J]. J of Intelligent & Fuzzy Systems, 2017, 32(4): 2875-2883.

[2] Ahmed M, Naser Mahmood A, Hu J. A survey of network anomaly detection techniques[J]. J of Network & Computer Applications, 2016, 60: 19-31.

[3] Erbacher R F, Walker K L, Frincke D A. Intrusion and misuse detection in large-scale systems[J]. Computer Graphics & Applications IEEE, 2002, 22(1): 38-47.

[4] Fathima S M H S S, Banu R S D W. Elliptical model for normal and abnormal gait classification [J]. Research J of Applied Sciences Engineering & Technology, 2015, 11(11): 1238-1244.

[5] Wang J, Wang S T, Deng Z H. Some problems in cluster analysis[J]. Control and Decision, 2012, 27(3): 321-328.

[6] Kang S H, Sandberg B, Yip A M. A regularized $k$-means and multiphase scale segmentation[J]. Inverse Problems & Imaging, 2017, 5(2): 407-429.

[7] Jeon Y, Yoo J, Lee J, et al. NC-Link: A new linkage method for efficient hierarchical clustering of large-scale data[J]. IEEE Access, 2017, 5: 5594-5608.

[8] Zhang X, Zhou A, Wang X, et al. Unmixing grain-size distributions in lake sediments: A new method of endmember modeling using hierarchical clustering[J]. Quaternary Research, 2017, 89(1): 1-9.

[9] Zhang Y, Geng G, Wei X, et al. Feature extraction of point clouds using the DBSCAN clustering[J]. J of Xidian University, 2017, 44(2): 114-120.

[10] Huang J, Hong Y, Zhao Z, et al. An energy-efficient multi-hop routing protocol based on grid clustering for wireless sensor networks [J]. Cluster Computing, 2017, 20(3): 1-13.

[11] Zhao Q H, Li X L, Zhao X M, et al. Fuzzy cluster image segmentation based on spatial constraint Student's-T hybrid model[J]. Control and Decision, 2016, 31(11): 2065-2070.

[12] Dai J, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification [J]. Applied Soft Computing J, 2013, 13(1): 211-221.

[13] Herawan T, Deris M M, Abawajy J H. A rough set approach for selecting clustering attribute[J]. Knowledge-Based Systems, 2010, 23(3): 220-231.

[14] Jiang Q, Huang B, Yan X. GMM and optimal principal components-based Bayesian method for multimode fault diagnosis [J]. Computers & Chemical Engineering, 2016, 84: 338-349.

[15] Chen S, Hong X, Harris C J. Probability density estimation with tunable kernels using orthogonal forward regression[J]. IEEE Trans on Systems, Man, & Cybernetics, Part B: Cybernetics, 2010, 40(4): 1101-1114.

[16] Laohakiat S, Phimoltares S, Lursinsap C. A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction[J]. Information Sciences, 2017, 381: 104-123.

[17] Jia R, Mao Z, Wang F. KPLS model based product quality control for batch processes[J]. Ciesc J, 2013, 64(4): 1332-1339.

[18] Xu W, Yan X, Xu W, et al. Application of single channel blind separation algorithm based on EEMD-PCA-robust ICA in bearing fault diagnosis[J]. Int J of Communications Network & System Sciences, 2017, 10(8): 138-147.

[19] Lahdhiri H, Elaissii I, Taouali O, et al. Nonlinear process monitoring based on new reduced rank-KPCA method[J]. Stochastic Environmental Research & Risk Assessment, 2017, 32(6): 1-16.

[20] Wu L Y, Li S L, Gan X S, et al. Network anomaly intrusion detection CVM model based on PLS feature extraction[J]. Control and Decision, 2017, 32(4): 755-758.

[21] Lin W C, Ke S W, Tsai C F. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors[J]. Knowledge-Based Systems, 2015, 78(1): 13-21.

[22] Tama B A. HFSTE: Hybrid feature selections and tree-based classifiers ensemble for intrusion detection system[J]. Ieice Trans on Information & Systems, 2017, 100(8): 1729-1737.

[23] Ashfaq R A R, Wang X Z, Huang J Z, et al. Fuzziness based semi-supervised learning approach for intrusion detection system[J]. Information Sciences, 2017, 378(C): 484-497.

[24] Lippmann R, Haines J W, Fried D J, et al. The 1999 DARPA off-line intrusion detection evaluation[J]. The Int J of Computer and Telecommunications Networking, 2000, 34(4): 579-595.


## 作者简介


刘金平 (1983–), 男, 副教授, 博士, 从事复杂工业过程自动化监控等研究, E-mail: ljp@hunnu.edu.cn;

张五霞 (1991–), 女, 硕士生, 从事计算机视觉与模式识别的研究, E-mail: wuxia@smail.hunnu.edu.cn;

唐朝晖 (1965–), 男, 教授, 博士, 从事复杂工业过程建模与故障诊断等研究, E-mail: zhtang@csu.edu.cn;

何捷舟 (1994–), 男, 硕士生, 从事计算机视觉与模式识别的研究, E-mail: hdc@smail.hunnu.edu.cn;

徐鹏飞 (1977–), 男, 副教授, 博士, 从事智能信息处理等研究, E-mail: xupf@hunnu.edu.cn.


(责任编辑: 李君玲)