# 基于免疫优势克隆网络聚类的入侵检测

白 琳
(西安邮电学院计算机学院 西安 710121)

**摘 要** 基于智能融合互补的观点,将免疫优势、倒位、克隆选择、非一致性变异和禁忌克隆等多种人工免疫系统算子引入网络结构聚类算法中,构造亲合度函数来指导聚类过程,得到一种能够自学习、自适应的进化网络来进行入侵检测数据的训练学习,通过该网络映射出大规模数据集的内在聚类结构,然后利用图论中的最小生成树对网络结构进行聚类分析,最终获得描述正常和异常行为的数据特征。在 KDD CUP99 数据集中进行了对比仿真实验,结果表明,该方法可高效地对大规模网络数据进行异常检测,以区分正常和攻击行为,并有效地检测出未知攻击。

**关键词** 免疫优势,非一致性变异,克隆选择,禁忌克隆,进化网络,入侵检测

**中图法分类号** TP393 **文献标识码** A

## Immunodominance-based Clonal Network Clustering Algorithm for Intrusion Detection

BAI Lin
(Dept. of Computer Science & Technology, Xi'an Institute of Post & Telecommunications, Xi'an 710121, China)

**Abstract** According to the idea of intelligent complementary fusion, a combination of immunodominance, inverse operation, clonal selection, non-uniform mutation and forbidden clone was employed in a novel clustering method with network structure for intrusion detection. The clustering process was adjusted in accordance with affinity function and evolution strategies. So an intelligent, self-adaptive and self-learning network was 'evolved' to reflect the distribution of original data. Then the minimal spanning tree was employed to perform clustering analysis and obtain the classification of normal and anormal data. The simulations through the KDD CUP99 dataset show that the novel method can deal with massive unlabeled data to distinguish normal case and anomaly and even can detect unknown intrusions effectively.

**Keywords** Immunodominance, Non-uniform mutation, Clonal selection, Forbidden clone, Evolutionary network, Intrusion detection

## 1 引言

入侵检测[1]被认为是继防火墙之后的第二道安全闸门,在不影响网络性能的情况下能对网络进行监测,从而提供对内、外部攻击和误操作的实时保护。入侵检测技术包括误用检测和异常检测[2]。误用检测建立攻击行为特征库,采用特征匹配的方法来确定攻击事件;异常检测建立用户正常行为模型,以是否显著偏离正常模型为依据进行检测,能够发现未知的攻击类型,是当前研究的热点。

现有的许多入侵检测系统对未知攻击的检测能力很有限。很多检测方法均采用有监督的学习算法[3],需要带标签的训练数据集来学习获得正常行为模型,如果训练数据的标签有错误,就会直接影响检测系统的有效性。况且收集大量带标签的数据集代价较高,不易实现。所以,本文采用无监督聚类学习算法来建立异常检测中的用户正常行为模型,并构建能够处理原始数据的入侵检测系统。

聚类的任务是把一个未标记的样本集按某种准则划分成若干子集,将相似的样本尽量归为一类,不相似的样本归为不同类[4]。这种方法可以定量地确定研究对象之间的亲疏关系,以达到合理的分类。本文以聚类技术为背景,提出一种基于免疫优势和克隆策略的进化网络聚类方法来构造入侵检测系统。

该方法借鉴了 Leandro 的进化人工免疫网络(Evolutionary Artificial Immune Network, AiNet)[5]思想,是一种通过抗体抗原间的相互作用、识别、记忆而进行进化学习的网络结构聚类方法:以抗体-抗原亲合度成熟为依据,通过对抗体群实施免疫优势、克隆增值、倒位操作、非一致性变异、克隆选择、克隆死亡以及禁忌克隆等进化操作,自动调整抗体群的规模和抗体对抗原的适应性,实现有方向的局部寻优和全局搜索。通过网络进化学习和自我抑制,经多次迭代找到满足要求的抗体群,得到聚类原型。

算法引入免疫优势算子,自适应动态地从抗体群本身获得先验知识来避免陷入局部最优,提高网络的进化速度;加入禁忌克隆算子,将克隆选择和禁忌克隆结合,使网络兼具免疫特异性和免疫耐受性,能够克服边界不清晰对聚类效果的影响;引入非一致性变异算子增强局部求解的自适应性、优化局

到稿日期:2011-09-02 返修日期:2011-11-09 本文受陕西省教育厅科研项目(2010JK837),西安邮电学院中青年科研项目(ZL2008-27)资助。

白 琳(1980—),女,硕士,讲师,主要研究方向为智能信息处理,E-mail:bailin@xupt.edu.cn。

部求解性能。

将这种基于免疫优势的克隆网络聚类算法用于入侵检测，首先用训练数据来“进化”一个抗体网络，完成数据压缩，使网络结构反映原始数据的分布情况，然后利用图论中的最小生成树对网络结构进行聚类分析，最终获得描述正常行为和异常行为的数据特征。在实际中，入侵检测数据具有混合属性特征，包括连续属性和离散属性。而现有的聚类算法在计算数据差异度时，常采用欧氏距离作为衡量标准，在处理网络数据时，通常只考虑其数值属性的值，而没有利用非数值属性的任何信息，这无疑会影响检测效果。因此本文采用混合型(离散的和连续的属性)数据差异度的衡量标准来提高聚类的精度[6]。

## 2 人工免疫系统

### 2.1 免疫优势

现代免疫学认为，抗原表面可以有一种或多种不同的抗原决定簇(表位)，每一种表位决定着相应的特异性。一个抗原分子上可有多个表位，但在诱导宿主免疫应答时，可能只有一种或一个表位起主要作用，使宿主产生以该特异性为主的免疫应答，这种现象称为免疫优势(Immunodominance)[7]，起关键作用的表位称为显性表位。免疫优势是在抗体与抗原相互作用中产生的。免疫优势位点决定了在自然选择中哪一种抗原将面临更大的压力。

#### 2.1.1 抗体免疫优势

抗体免疫优势表示抗体编码每一位对抗原的不同重要程度。对于抗体 $a=(a_1,a_2,\cdots,a_l)$，抗体 $a$ 的解码为 $e^{-1}(a)$，如果对于任意的 $a_k\in\{0,1\}$，$k\neq i$ 都有 $f(a)=f(e^{-1}(a_1,a_2,\cdots,a_{i-1},1,a_{i+1},a_l))\geqslant f(e^{-1}(a_1,a_2,\cdots,a_{i-1},0,a_{i+1},a_l))$或 $f(a)=f(e^{-1}(a_1,a_2,\cdots,a_{i-1},0,a_{i+1},a_l))\geqslant f(e^{-1}(a_1,a_2,\cdots,a_{i-1},1,a_{i+1},a_l))$，则称抗体编码的第 $i$ 位具有免疫优势，该位的取值称为优势值。抗体具有免疫优势的过程称为获得免疫优势。抗体免疫优势的本质是将先验知识引入算法，以改善算法性能。

### 2.2 克隆选择算法

克隆选择算法[8]包括 3 个步骤：克隆、免疫基因操作和克隆选择。克隆的实质就是在进化过程中，在每一代最优解的附近，根据亲合度的大小进行克隆，产生一个变异解的群体，增加抗体的多样性，扩大了搜索范围，并防止了进化早熟和搜索陷于局部最优[9]。

### 2.3 进化人工免疫网络

Leandro 根据 Jerne 的免疫网络理论，提出一种进化人工免疫网络，其主要思想是：设 $X=\{x_1,x_2,\cdots,x_n\}$ 是待聚类对象的全体，其中 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})'$ 表示第 $i$ 个样本的 $p$ 个特征值，$x_i$ 可用状态空间 $S^m$ 中的一个点 $s$ 来表示，将 $s$ 作为抗原，来决定抗体-抗体以及抗体-抗原之间的相互作用。且系统内部的相互作用可以用一个连通图来表示。

网络模型定义为：免疫网络是一个加权的图 $G$，该图由一组不完全连接的神经元节点构成，每对节点产生一条边，边的长度称为权值或连接强度。

免疫网络具有数据压缩功能，可以解决聚类前需要类数、依赖先验知识的问题。但是当数据集中各类的类边界不明显或存在噪声时，这些样本作为抗原能够极大地激活免疫系统，引起细胞增殖和抗体分泌，使网络结构不够清晰，影响分类效果。

## 3 基于免疫优势的克隆网络聚类算法

### 3.1 距离测度

采用下列混合属性相异测度函数：

$$d(x_i,x_j)=\sqrt{\sum_{k=c_1}^{c_{m_c}}|x_{ik}-x_{jk}|^2+\lambda\cdot\sum_{l=d_1}^{d_{m_d}}\delta(x_{il},x_{jl})} \tag{1}$$

式中，根号内第一项是连续属性相异度；第二项是离散属性相异匹配测度；常数 $\lambda$ 用来调节两种属性在目标函数中的比例；$\delta(\cdot)$ 表示离散属性相异程度，定义为：

$$\delta(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases}$$

### 3.2 亲合度函数

根据类的目标函数 $C(W,P)$，聚类的目的是要获得数据集 $X$ 的模糊划分矩阵 $W$ 和聚类原型 $P$。$W$ 和 $P$ 是相关的，已知其一即可求得另一个的解，可令一组聚类原型 $P$ 就是一个抗体。根据目标函数越小，聚类效果越好，抗体-抗原亲合度越大的原则来构造抗原 $x_i$ 与抗体 $p_j$ 的亲合度函数 $f(x_i,p_j)$，$f$ 越大，$p_j$ 越接近 $x_i$。根据距离测度函数构造 $f$ 如下：

$$f(x_i,p_j)=\frac{N}{1+\sum_j|p_j-x_i|^2+\lambda\cdot\sum_j\delta(p_j,x_i)} \tag{2}$$

式中，$N$ 是与数据预处理相关的常量，由实验测试值确定。抗体-抗体亲合力 $s_{ij}$ 等于它们的距离测度 $s_{ij}=d(p_i,p_j)$。

### 3.3 抗体免疫优势获得

抗体免疫优势算子构造如下：采用二进制编码的抗体群落，记抗体群 $A=[a_1\quad a_2\quad a_N]$为：

$$A=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1l}\\ a_{21} & a_{22} & \cdots & a_{2l}\\ \vdots & \vdots & \ddots & \vdots\\ a_{N1} & a_{N2} & \cdots & a_{Nl}\end{bmatrix} \tag{3}$$

定义参考抗体 $m=[m_1\quad m_2\quad\cdots\quad m_l]$，其中：

$$m_i=\begin{cases}1, & \frac{1}{N}\sum_{j=1}^{N}a_{ji}\geqslant\frac{1}{2}\\ 0, & \text{others}\end{cases} \tag{4}$$

设 $a_i\in A$ 是抗体群中的最好抗体，根据式(3)、式(4)如果 $f(m)>f(a_i)$，则互换。

以一定的概率 $p_{id}$，使 $a_j\in A$，$j=1,2,\cdots,l$；and $j\neq i$ 获得免疫优势，具体地，令：$a_j=H(a_j+a_i-m-1)$；$H(\cdot)$是一个门限函数：$H(x)=\begin{cases}1, & x>0\\ 0, & \text{others}\end{cases}$，$p_{id}$是一个自适应调节的参数，如果 $f(a_i)>f(a_j)$或 $f(m)>f(a_i)$，表明免疫优势的作用是有效的，则增大 $p_{id}$，否则减小。

### 3.4 非一致性变异

如果 $y=(v_1,\cdots,v_k,\cdots,v_n)$是一个父解，分量 $v_k$ 被选中进行变异，其定义区间为$[a_k,b_k]$，则变异后的解为 $y'=(v_1,\cdots,v_{k-1},v_k',\cdots,v_n)$，其中：

$$v_k'=\begin{cases}v_k+\Delta(i_t,b_k-v_k), & \text{如果 } rnd(2)=0\\ v_k-\Delta(i_t,v_k-a_k), & \text{如果 } rnd(2)=1\end{cases} \tag{5}$$

式中，$rnd(2)=0$ 表示随机均匀地产生的正整数模 2 时所得的结果。$i_t$ 为当代演化代数，函数 $\Delta(i_t,s)$的值域为$[0,s]$，并使得当 $i_t$ 增大时，$\Delta(i_t,s)$接近于 0 的概率增加。$\Delta(i_t,s)=s$

$\cdot(1-r^{(1-i_t/T)^{\lambda}})$，其中 $r$ 为[0,1]内的随机数，$T$ 为最大进化代数，$\lambda$ 为决定非一致性程度的参数，起着调整局部搜索区域的作用，一般取 2～5，本文取 3。

### 3.5 基于免疫优势的克隆网络学习算法

Step 1 $l=1$，初始化抗体群 $A$：随机生成初始网络，网络节点即抗体 $y_j(j=1,2,\cdots,N_A)$ 是 $N_A$ 个 $p$ 维向量，设置各项参数和终止条件。

Step 2 计算当前抗体群中各抗体间的亲合力。

Step 3 获得免疫优势：$l=l+1$，对抗体亲合力排序，选取当前抗体群 $A(l)$ 中具有较大亲合力的、比例为 $\beta_1$ 的优秀抗体作为提取免疫优势的抗体源，进行编码，按 3.3 节相关操作，生成参考抗体 $m$，将抗体中所具有的免疫优势依概率 $p_{id}$ 推广到其它抗体中，得到抗体群 $A_A(l)$。

Step 4 对每个输入的抗原 $x_i(i=1,\cdots,n)$，对 $A_A(l)$ 做如下操作：

Step 4.1 按式(2)计算 $A_A(l)$ 中抗体-抗原的亲合度；

Step 4.2 倒位操作：按概率 $p_d$ 选取 $A_A(l)$ 中抗体进行倒位操作，设倒位点为 $p,q,p<q$，则对抗体 $A_i=\{x_{i1},\cdots,x_{ip},\cdots,x_{iq},\cdots,x_{im}\}$ 的倒位操作为：

$$A_i'=\{x_{i1},x_{i2},\cdots,x_{iq},x_{i(q-1)},\cdots,x_{i(p+1)},x_{ip},\cdots,x_{im}\} \quad (6)$$

Step 4.3 对倒位后的抗体计算抗体-抗原的亲合度；

Step 4.4 按比例选择 $k$ 个亲合度最高的抗体($y_{r_1},\cdots,y_{r_k}$)，按亲合度越高，克隆规模越大的原则进行克隆：

$$T_c^C(y_{r_m})=I_m\times y_{r_m},m=1,2,\cdots,k \quad (7)$$

式中，$I_m$ 为元素为 1 的 $q_m$ 维行向量。

$$q_m=Int(n_c*\frac{f(x_i,y_{r_m})}{\sum_{b=1}^{k}f(x_i,y_{r_m})}),m=1,\cdots,k \quad (8)$$

式中，$Int(c)$ 表示大于 $c$ 的最小整数，$n_c$ 为克隆总规模。

Step 4.5 非一致性变异：对克隆后的抗体，根据概率 $P_m$ 产生父代，指定个体编码的每个基因作为变异点进行 3.4 节的变异操作；

Step 4.6 计算抗体-抗原亲合度，选取 $\varepsilon\%$ 亲合度最高抗体组成记忆矩阵 $M_A$；

Step 4.7 克隆死亡：删除 $M_A$ 中节点亲合度小于亲合度门限 $\sigma_k$ 的神经元来减小矩阵规模；

Step 4.8 计算 $M_A$ 中抗体-抗体亲合力 $s_{ij}'$，使 $s_{ij}'$ 小于压缩门限 $\sigma_s$ 的抗体死亡；

Step 4.9 将 $M_A$ 和原始抗体 $A$ 结合：$A\leftarrow[A,M_A]$。

Step 5 禁忌克隆：对 $A$ 中任一抗体，计算与该抗体亲合力小于门限 $\sigma_s$ 的抗体个数 $n$，若 $n$ 小于 $\sigma_f$，将该抗体删除。

Step 6 网络压缩：计算 $A$ 中抗体-抗体亲合力 $S_{ij}$，使 $S_{ij}<\sigma_s$ 的抗体死亡，随机选若干抗体代替已删抗体加入 $A$。

Step 7 若满足终止条件(当前最优抗体连续 10 代无改进)，则转 Step 8，否则对抗体群中比例为 $\beta_2$ 的具有最小亲合度的抗体，重新初始化，返回 Step 2。

Step 8 网络输出，算法停止。

算法分析：

(1)在获得抗体免疫优势的过程中通过引入先验知识和在线自适应动态获得先验知识的机制，实现了个体间的信息交换，更好地保证了抗体种群的多样性；利用搜索过程中获得的问题本身的先验知识，有效地避免了陷入局部极小值的问题，提高了算法性能。

(2)倒位操作增加了抗体群中个体的多样性。

(3)非一致性变异提高了算法的局部搜索能力，有利于所得解的多样性。在进化前期，解在较大的邻域半径内搜索，随着迭代次数不断增大，解的搜索邻域半径也逐渐变小，体现了搜索的局部求解自适应性能力。

(4)克隆死亡和网络压缩将网络规模动态调整到合理范围。

(5)对亲合度高的个体进行克隆，将全局和局部搜索结合，避免陷入局部极值的问题；通过禁忌克隆操作来消除稀疏抗原对网络的刺激，使网络产生免疫耐受性。因此，该网络兼具免疫特异性和免疫耐受性。同时，算法在每代个体中进行高频变异和克隆选择，使网络的动态性能和结构受到进化策略的控制，所以该网络是“进化的”。

(6)网络的输出是由代表抗原网络内部图像的记忆矩阵和决定网络节点间联系并描述网络结构的内部亲合矩阵组成。为了确定数据集的聚类结构，采用 3.6 节方法分析。

### 3.6 聚类分析

由 3.5 节聚类算法得到的网络 $A$ 是聚类样本 $X$ 的空间特征浓缩，那么对 $X$ 的聚类分析即转换为对记忆网络 $A$ 的聚类分析，本文采用简化的连通图最小生成树的方法来实现[10]。

**定义** 连通图 $G$ 的一个子图如果是一棵包含 $G$ 所有节点的树，那么该子图就是 $G$ 的生成树；使各边权值之和最小的生成树是 $G$ 的最小生成树。

聚类分析算法：

Step 1 对输出的网络 $A$ 构建最小生成树 MST；

Step 2 剪枝：剪断 MST 中权值大于剪枝系数 $\sigma_m$ 的边；

Step 3 抗体聚类：剪枝后的 MST 即由多棵相互断开的子树构成，组成每棵子树的抗体即形成一个抗体聚簇；

Step 4 用条形(Bar)图描述 MST 的边长，Bar 图中山谷的数目即为数据集的聚类类别数。

## 4 基于免疫优势克隆网络的入侵检测

### 4.1 正常模型确定

通过 3.5 节和 3.6 节算法聚类分析得到聚类原型，抗体群被划分为 $h$ 个子集，每个子集 $P_c(1\leqslant c\leqslant h)$ 中包含若干个抗体。每个子集内的抗体互相接近，与其它子集中的抗体则相距较远。给每个子集设定标签。

抗体的分类映射了训练数据的分类，对训练集 $X=\{x_1,\cdots,x_n\}$，计算 $x_i(1\leqslant i\leqslant n)$ 和各抗体子集的距离，找到最短距离 $d(x_i,P_m)$ $(1\leqslant m\leqslant h)$，那么第 $m$ 个抗体子集的标签就是抗原 $x_i$ 所属的类别。据此完成对训练数据的分类。

对分好类的训练数据子集需要确定正常类和异常类。检测系统基于两个合理的假设[5]：同类数据在合理的尺度条件下，在特征空间中互相接近，不同类数据彼此远离；入侵行为与正常行为本质特性差异很大，且相对很少，这在实际中是合理的。因此，根据样本分布，就可以将子类中的数据量划分为各个正常类和异常类。如果某类的数据量与样本数据总量之比不小于 $r(0<r<1)$，将其标记为正常。对应于正常类的抗体是正常数据的代表点，是正常行为模型。

### 4.2 检测算法

根据正常和异常模型进行异常检测：

Step 1　对数据集 $X=\{x_1,\cdots,x_n\}$，计算 $x_i(1\leqslant i\leqslant n)$和各原型 $p_j$ 的距离，找到最短距离 $d_{\min}(x_i,p_{\min})(1\leqslant \min\leqslant N_A)$。

Step 2　$p_{\min}$就是 $x_i$ 的代表点，根据 $p_m$ 所属模型的类别来确定 $x_i$ 是否异常。

Step 3　若 $d_{\min}(x_i,p_m)\geqslant\xi$，则判断 $x_i$ 为未知攻击。

## 5　仿真实验

### 5.1　实验数据

实验采用 KDD CUP99[11] 入侵检测数据集，该数据集共计 4,900,000 多条连接记录，包含 9 个星期的网络流量，其中 7 周时间的训练数据包含约 500 万条网络连接记录，2 周时间的测试数据约包含 200 万条连接记录。每条记录都带有持续时间、协议类型、传输的字节数等参数，参数属性共计 42 个，分为 33 个连续属性和 8 个离散属性，包括有字符枚举型、连续自然数型、连续整数型、标签等。

KDD CUP99 数据集包含 4 大类攻击：拒绝服务攻击(DOS)；对本地超级用户的非法访问(U2R)；未经授权的远程访问(R2L)；扫描与探查(Probing)。在 KDD CUP99 的训练数据中，共有 23 个不同的连接标志，除“正常”外，其余 22 个代表攻击类型，这些攻击可以分成如前所述的 4 类攻击模式。测试集中共包含 38 个不同的连接标志，除“正常”外，其余 37 种代表攻击类型。在 37 种攻击类型中，有 15 种未在训练集中出现，对训练数据而言是新的攻击模式，属于未知攻击。

我们从 KDD CUP99 训练集中随机选取 15 万条记录作为训练样本，其中入侵记录为 1600 条(包含 4 大类 22 种攻击类型)，包括 DOS 攻击 750 条、U2R 攻击 160 条、R2L 攻击 360 条、Probing 攻击 330 条。由于训练集仅含有 22 种类型的攻击，而 KDD CUP99 测试集总共包含 37 种攻击类型，则对于训练集而言，存在 15 种未知攻击。为了测试算法对未知攻击的检测效果，从 KDD CUP99 测试集中各选取 10 万条数据作为两个测试集，其中入侵数据为 1100 条(包含未知攻击和已知攻击类型在内的 4 大类 37 种攻击类型)。

### 5.2　数据预处理

将数据集中字符枚举特征变成离散数值特征，用不同的自然数分别代表不同的字符枚举型属性值。例如，用 0 表示“http”协议，1 表示“ftp”协议。

数据离散化：离散化的任务是把连续属性的取值范围或取值区间划分为若干个数目不太多的小区间，每个小区间对应一个离散的符号。本文采用等分区间法[12]实现离散化。

### 5.3　算法参数设置

3.5 节算法中克隆规模系数 $n_c=120$；获得免疫优势的概率 $p_{id}=0.15$；克隆选择多样性控制参数为 0.15，此值越大，选择后的抗体群多样性就越好；用于提取免疫优势的优秀抗体在抗体群中的比例 $\beta_1=0.07$；连续 5 代无改进时抗体群中较差抗体更新比例 $\beta_2=0.2$；变异概率 $p_m=0.25$；网络压缩门限 $\sigma_s$ 控制着抗体间的亲合度，通过调整网络中抗体细胞的特异性水平来控制网络的可塑性及聚类精度，对聚类结果和性能以及网络规模影响很大，通常先设较小值($10^{-3}$)，再根据实验结果不断增加，最终确定到合适的值，使网络具有较好的性能，从而得到合理的聚类结果。本文最终确定 $\sigma_s=0.25$。3.6 节剪枝系数 $\sigma_m$ 取 3.5。参数取值主要根据实验经验值得到。

### 5.4　实验结果

利用训练数据进化出免疫优势克隆网络，再用 MST 剪枝算法对网络结构进行分析，得到 Bar 图，见图 1。

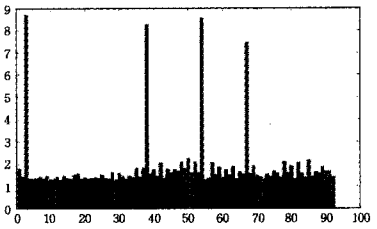


图 1　本文算法的 Bar 图

根据图中山谷数目可知聚类类别数为 5。

将文献[5]免疫网络算法和文献[13]自适应多克隆聚类算法以及文献[14]方法作为本文方法的比较算法。文献[13]所提算法是一种采用多克隆算子的网络结构聚类方法，该算法的思想和本文类似；文献[14]所提算法是一种简单的基于距离的聚类方法，其首先将某个训练数据作为第一类的中心，然后对每个数据，从当前所有聚类中心找到距离该数点最近的中心，若此最短距离小于预设的聚类半径，则将该点归入这一聚类中心所在的类，否则将该点作为一个新类的中心。

经过 20 次独立实验，在平均的情况下，4 种算法的聚类结果见表 1—表 3。对于聚类结果，类间距大、类内距离小，算法性能就高[15]。本文算法满足这一指标并优于其它算法。

表 1　各聚类算法类间距

| 算法 | 类别 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 本文 | 1 | 9.12 | 9.01 | 8.97 | 9.06 | \ |
| | 2 | \ | 8.76 | 8.58 | 8.56 | \ |
| | 3 | \ | \ | 8.71 | 8.93 | \ |
| | 4 | \ | \ | \ | 8.75 | \ |
| 文献[5] | 1 | 7.32 | 7.56 | 7.35 | 7.21 | \ |
| | 2 | \ | 7.03 | 6.77 | 7.27 | \ |
| | 3 | \ | \ | 6.33 | 7.06 | \ |
| | 4 | \ | \ | \ | 6.25 | \ |
| 文献[13] | 1 | 8.06 | 8.01 | 7.78 | 8.16 | 7.91 |
| | 2 | \ | 7.81 | 7.89 | 8.07 | 7.76 |
| | 3 | \ | \ | 7.57 | 7.29 | 7.29 |
| | 4 | \ | \ | \ | 7.87 | 7.56 |
| | 5 | \ | \ | \ | \ | 7.12 |
| 文献[14] | 1 | 5.67 | 5.25 | 5.85 | 5.16 | \ |
| | 2 | \ | 5.12 | 5.68 | 6.09 | \ |
| | 3 | \ | \ | 5.13 | 5.36 | \ |
| | 4 | \ | \ | \ | 5.67 | \ |

表 2　各聚类算法类内距

| 算法 | 类别 | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 本文 | 1.27 | 1.12 | 1.32 | 1.38 | 1.47 | \ |
| 文献[5] | 2.18 | 2.87 | 2.28 | 2.59 | 2.71 | \ |
| 文献[13] | 1.83 | 1.97 | 1.86 | 1.65 | 1.73 | 1.62 |
| 文献[14] | 3.51 | 2.92 | 3.12 | 3.29 | 2.98 | \ |

表 3　各算法标类结果(N-正常类，A-异常类)

| 算法 | 类别 | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 本文 | N | A | A | A | A | \ |
| 文献[5] | N | A | A | A | A | \ |
| 文献[13] | N | N | A | A | A | A |
| 文献[14] | N | N | A | A | A | \ |

类间距定义为：

$$D(C_i,C_j)=\frac{1}{n_i n_j}\sum_{p\in C_i}\sum_{q\in C_j}d(p,q) \quad (9)$$

式中，$n$ 为各类所含对象数。

类内距定义为：

$$D_c=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}d(x_i,x_j),c\in[1,r] \quad (10)$$

式中，$r$ 为聚类数。

入侵检测相关的统计量定义为[3]：检测率是指被检测出来的异常样本数占异常样本总数的百分比；误警率是指正常样本被识别成异常样本的数目占正常样本总数的百分比。

通过聚类得到正常模型后，即可进行异常检测。本文算法将训练集分为 5 类，那么对于某测试数据，当它属于第 1 类时，将其标为正常，否则，不管它属于何种攻击类，都将其作为异常数据来计算总检测率（此数据应该是被正确检测到的入侵数据）；对于正常数据，无论将其误认到哪个攻击类中，都将其作为误警数据来计算误警率。表 4 给出了 4 种算法经 20 次独立实验的平均检测结果（已知入侵指训练集中包含的 22 种攻击；未知入侵指训练集中未包含的 15 种攻击）；表 5 给出了本文算法对 4 大类攻击的检测结果。

表 4　各算法检测结果

| 数据集 | 算法 | 检测率 | | |
|---|---|---|---|---|
| | | 已知 | 未知 | 误警率 |
| 测试集 1 | 本文 | 91.03% | 82.92% | 3.87% |
| | 文献[5] | 81.01% | 72.16% | 6.37% |
| | 文献[13] | 87.96% | 81.12% | 4.65% |
| | 文献[14] | 78.92% | 57.21% | 8.25% |
| 测试集 2 | 本文 | 90.78% | 83.45% | 4.01% |
| | 文献[5] | 82.09% | 69.98% | 7.17% |
| | 文献[13] | 88.79% | 79.28% | 5.11% |
| | 文献[14] | 77.12% | 61.23% | 8.56% |

表 5　本文算法对各类攻击检测效果

| 数据集 / 攻击类型 | 测试集 1 | | 测试集 2 | |
|---|---|---|---|---|
| | 检测率 | | 检测率 | |
| | 已知 | 未知 | 已知 | 未知 |
| DOS | 86.91% | 77.67% | 83.12% | 80.09% |
| U2R | 96.16% | 88.96% | 96.87% | 91.65% |
| R2L | 83.90% | 71.21% | 85.02% | 69.09% |
| Probing | 97.13% | 93.82% | 98.01% | 92.99% |

从表 1－表 3 聚类结果来看，本文算法聚类效果是优秀的，且对训练集的分类是有效的。具体分析如下：

免疫优势操作自适应动态地利用了抗体群自身的先验知识，有效地避免了早熟和陷入局部极小值的问题；克隆选择算法能以概率 1 收敛到全局最优解，克隆对应着一个亲合度成熟的过程，即对亲合度较低的个体在克隆选择机制作用下，经历增殖和变异操作后，其亲合度逐步提高而“成熟”，因此克隆策略控制着网络的抗体多样性、特异性和进化速度；非一致性变异体现了搜索的局部求解自适应性能力，保留了最佳个体并改进较差个体，和倒位算子、克隆策略共同导致算法在保证候选解多样性的前提下，将全局和局部搜索更加有机地结合，使本文聚类算法拥有更好的聚类性能（类间距大而类内聚小）。入侵检测的检测效果十分依赖正常模型的建立，正是这些良好的聚类性能使得正常模型的有效性和可靠性较高。在距离尺度下，入侵检测数据边界并不是很清晰，特别是如果网络学习时带有随机性，有可能会进一步模糊数据边界；本文算法引入禁忌克隆操作来克服边界不清晰对聚类效果的影响，将克隆选择和禁忌克隆相结合，使进化网络兼具免疫特异性和免疫耐受性。

从表 4 的检测结果来看，本文算法的检测率和误警率都优于其他 3 种方法，分析如下：

(1)实验训练样本达到 15 万条，测试样本达到 10 万条，对已知入侵的检测率达到 90%以上，说明本文算法对处理大规模数据是有效的。

(2)文献[5,13]的算法和本文算法都是基于抗体抗原相互作用的进化网络结构聚类，其中文献[13]采用多克隆算子对抗体进行重组、变异和选择操作，提高了局部搜索和脱离局部极小值的能力，多克隆算子的子代更多地继承了父代的特点，虽然其收敛速度较快，但是多样性有所下降，使网络学习和进化受到影响。并且前两种方法对边界模糊问题处理能力有限。而本文算法引入了免疫优势、禁忌克隆等操作，整体性能较优，尤其是禁忌克隆操作有效地解决了边界模糊问题。

(3)文献[14]的算法是一种简单的基于距离的无监督方法，其没有采用进化算子，所以检测效果比其他 3 种方法低；但计算速度最快、复杂度较低，对每个数据只扫描一遍，简单、快速。由于它对每一类只保留一个点，而本文方法对每一类保留若干个点，因此本文算法能更好地逼近任意形状分布的数据，与数据分布无关。

从表 5 的检测结果来看，本文算法对 U2R 和 Probing 攻击检测效果理想，但其他两种效果一般，这种结果也符合我们的预估：由于和其它聚类方法一样，本文算法也是基于距离的，因此对某些伪装合法用户的攻击（某些 DOS 和 R2U 攻击）检测率不高，这是因为它们的特征与正常数据很相似，在距离测度下不易与正常数据区分，从而增加了检测的难度。

**结束语**　本文构造了一种基于免疫优势和克隆策略的网络结构聚类算法用于入侵检测，新的聚类算法以亲合度成熟为依据，结合多种进化操作，具有自学习、自适应和快速收敛的性能。基于该方法的入侵检测系统能够有效处理大规模的、原始的、具有混合属性的网络数据，得到较好的检测结果，并能有效地检测到未知攻击；但本文聚类算法的参数较多，实验中使用试探法根据经验值选取，如何自适应地设定相关参数来提高算法性能将是一个有待研究和解决的问题。

## 参 考 文 献


[1] 戴英侠，连一峰，王航. 系统安全与入侵检测[M]. 北京：清华大学出版社，2002

[2] 蒋建春，马恒太，任党恩. 网络安全入侵检测：研究综述[J]. 软件学报，2000，11(11)：1460-1466

[3] Portnoy L. Intrusion Detection with Unlabeled Data using Clustering[D]. Columbia University，December，2000

[4] Everitt B. Cluster Analysis[M]. Heinemann Educational. Books Ltd，1974

[5] de Castro L N，Von Zuben F J. An Evolutionary Immune Network for Data Clustering [C]// Proc. of the IEEE SBRN. 2000：84-89

[6] 李洁，高新波，焦李成. 基于克隆算法的网络结构聚类新算法[J]. 电子学报，2004，32(3)：1195-1199

[7] Liu Ruo-chen，Shen Zheng-chun，Jiao Li-cheng，et al. Immunodomaince Based Clonal Selection Clustering Algorithm [C]// Proceedings of The 2010 IEEE Congress on Evolutionary Computation，CEC2010. Barcelona，Spain，2010：18-23


（下转第 118 页）

无人机现有的 FCS 系统展开研究，通过分析其设计文档及源代码，采用 MARTE 建立起 FCS 系统的软件模型，并给出了基于时间自动机的系统动态行为的形式化模型实例；结合无人机 FCS 系统的应用背景，给出了基于时间自动机模型的测试用例生成方法，包括建立测试用例生成框架、测试用例生成规则以及用例生成策略等。然后，对某型无人机 FCS 系统中主控模块进行建模与测试用例生成的实例分析研究。实例模型的应用分析表明，所提基于时间自动机模型的测试用例集生成方法是可行的，对于提高现有及未来无人机 FCS 系统的测试的有效性具有很好的指导意义。

目前正在继续开展的工作包括：研究基于更多覆盖准则（如数据流覆盖等）的测试用例生成方法，并且对某型无人机飞控软件的其他模块，如控制律解算模块、导航模块、地面检测模块、遥控遥测模块等进行建模和测试用例生成，逐步构建一个完整的无人机 FCS 系统的测试生成、检查及维护的方案，全面提高 FCS 系统测试用例生成的有效性以及系统的可维护性；同时为进一步建立与完善无人机 FCS 系统的各类模型，深入展开无人机应用领域中基于模型的分析与检验方法奠定良好的基础。

## 参 考 文 献


[1] 杨一波，陈欣. 无人机飞行控制软件测试技术研究[D]. 南京：南京航空航天大学，2008

[2] 王鑫，陈欣，张民. 基于 SCADE 的无人机飞行控制系统软件设计[D]. 南京：南京航空航天大学，2008

[3] Mussa M, Ouchani S, Sammane W A, et al. A Survey of Model-driven Testing Techniques[C]// 2009 Ninth International Conference on Quality Software. 2009: 167-172

[4] Heckel R, Lohmann M. Towards Model-Driven Testing[J]. Electronic Notes in Theoretical Computer Science, 2003, 82(6): 33-43

[5] 张天，张岩，于笑丰，等. 基于 MDA 的设计模式建模与模型转换[J]. 软件学报，2008，19(9)：2203-2217

[6] Object Management Group, OMG[OL]. http://www.omg.org/

[7] 张天，李宜东，Jouault F，等. 基于 MDE 的异构模型转换：从 MARTE 模型到 FIACRE 模型[J]. 软件学报，2009：214-233

[8] OMG. UML Profile for MARTE, Beta 2[EB/OL]. http://www.omg.org/cgi-bin/doc? ptc/2008-06-08, 2008

[9] Clarke E M, Wing J M. Formal methods: state of the art and future directions[J]. ACM Computing Surveys, 1996, 28(4): 626-643

[10] Herber P, Fellmuth J, Glesner S. Model checking SystemC designs using timed automata[C]// Proceedings of the 6th IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis. 2008: 131-136

[11] Lampka K, Perathoner S, Thiele L. Analytic Real-time Analysis and Timed Automata: A Hybrid Methodology for the Performance Analysis of Embedded Real-time Systems[J]. Design Automation for embedded systems, 2010: 193-227

[12] Behrmann G, David A, Larsen K G. A Tutorial on Uppaal[Z]. Formal Methods for the Design of Real-Time Systems, 2004

[13] Hessel A. Model-Based Test Case Generation for Real-time Systems[R]. 1214.51. Faculty of Science and Technology, 2007

[14] 颜炯，王戟，陈火旺. 基于模型的软件测试综述[J]. 计算机科学，2004，27：184-187

[15] Gutierrez J J, Escalona M J, Mejias M, et al. An approach for Model-driven test generation[C]// Research Challenges in Information Science, 2009. RCIS2009. Third International Conference on. April 2009: 303-312

[16] Uppaal CoVer[EB/OL]. http://www.uppaal.org/cover/, 2005

[17] Hessel A, Pettersson P. CoVer-A Test Case Generation Tool for Real-time Systems[C]// FATES07. Tallinn, Estonia, 2007

[18] Nielsen B, Skou A. Automated test generation from timed automata [J]. International Journal on Software Tools for Technology Transfer, 2003(5): 59-77

[19] Hessel A, Larsen K G, Nielsen B, et al. Time-optimal real-time test case generation using Uppaal[C]// Lecture Notes in Computer Science. 2004: 136-151

[20] Myers G. The Art of Software Testing[M]. Wiley-Interscience, 1979

[21] Hessel A, Larsen K G, Nielsen B, et al. Testing real-time systems using Uppaal[C]// Formal Methods and Testing, LNCS 4949. 2008: 77-117

[22] Hessel A, Pettersson P. A global algorithm for model-based test suite generation[C]// Third Workshop on Model-Based Testing. Braga, Portugal, 2007

[23] 王鑫，陈欣，张民. 基于 SCADE 的无人机飞行控制系统软件设计[D]. 南京：南京航空航天大学，2008

[24] Yin Yong-feng, Li Zhen, Liu Bin. Real-time Embedded Software Test Case Generation Based on Time-extended EFSM: A Case Study[C]// WASE International Conference on Information Engineering. 2010: 272-275


（上接第 86 页）


[8] 焦李成，杜海峰. 人工免疫系统进展与展望[J]. 电子学报，2003，31(10)：1540-1549

[9] 杨咚咚，焦李成，公茂果，等. 求解偏好多目标优化的克隆选择算法[J]. 软件学报，2010，21(1)：14-33

[10] Leclerc B. Minimum spanning trees for tree metrics: abridgements and adjustments[J]. Journal of Classification, 1995, 12: 207-241

[11] kdd cup99 dataset[OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999

[12] Catlett J. On changing continuous attributes into ordered discrete attributes [C]// EWSL' 91. 1991: 164-178

[13] Ma Li, Jiao Li-cheng, Bai Lin, et al. Polyclonal Clustering Algorithm and its Convergence[J]. The Journal of China Universities of Posts and Telecommunications, 2008, 15(3): 110-117

[14] 罗敏，王丽娜，张焕国. 基于无监督聚类的入侵检测方法[J]. 电子学报，2003，11(11)：1714-1716

[15] 杨草原，刘大有，杨博，等. 聚类集成方法研究[J]. 计算机科学，2011，38(2)：166-170