基于激素调节免疫网络聚类的入侵检测系统

白 琳[1], 杨 超[2]

[1]西安邮电大学 计算机学院 西安 中国 710121
[2]西安电子科技大学 网络与信息安全学院 西安 中国 710071

摘要 生物体的内分泌系统是一个高度进化的智能系统, 通过激素调节着生物体的神经、免疫系统。受其启发而得到的人工内分泌系统具有强大的调控机制, 将其内分泌激素用来调节人工免疫网络的抗体种群进化过程, 利用亲合度函数动态调节抗体的克隆规模和网络压缩的规模, 充分发挥优秀个体的先进特性来刺激亲合度成熟, 并能动态调控种群规模, 实现自适应、智能化的网络学习, 尤其当样本集边界模糊以及存在噪声样本时, 该网络依然可以通过自适应调节有效聚类。最终进化出一个小规模网络来映射原始入侵检测数据集的内在结构。最后, 利用图论中的最小生成树对网络结构进行分析, 获得描述正常和异常行为的数据特征, 得到入侵检测系统的正常模型, 由此构建出入侵检测系统。通过在 KDD CUP 数据集的对比仿真实验, 验证了该系统的有效性和可行性, 以及对未知攻击的检测能力。

关键词 人工内分泌系统; 激素; 免疫网络; 聚类分析; 入侵检测
中图法分类号 TP393 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2019.09.03

Intrusion Detection System Based on Hormone-Regulated Immune Network Clustering

BAI Lin [1], YANG Chao[2]

[1] School of Computer Science and Technology, Xi'an University of Post and Telecommunications, Xi'an 710121, China
[2]School of Cyber Engineering, Xidian University, Xi'an 710071, China

Abstract The endocrine system of organisms is a highly evolved intelligent system, which regulates the nervous and immune systems of organisms through hormones. The artificial endocrine system inspired by it has powerful regulation mechanism. Its endocrine hormones are used to regulate the evolution process of antibody population in artificial immune network. Affinity function is used to dynamically adjust the clone size and network compression size of antibody, and give full play to the advanced characteristics of excellent individuals to stimulate affinity maturity, and dynamically adjust the population size to achieve self-adaptation. The adaptive and intelligent network learning, especially when the boundary of the sample set is blurred and there are noisy samples, can still be effectively clustered by adaptive adjustment. Finally, a small-scale network was evolved to map the intrinsic structure of the original intrusion detection data set. Finally, the minimum spanning tree in graph theory is used to analyze the network structure, and the data characteristics describing normal and abnormal behavior are obtained, and the normal model of intrusion detection system is obtained, thus the intrusion detection system is constructed. The validity and feasibility of the system and the ability to detect unknown attacks are verified by the comparative simulation experiments on KDD CUP datasets.

Key words artificial endocrine system; hormone regulation; immune network; cluster analysis; intrusion detection

1 概述

入侵检测系统能够在网络系统受到危害之前拦截和响应入侵, 有效为网络信息安全系统保驾护航, 是一种非常重要和实用的实时网络安全技术。

入侵检测技术包括误用检测和异常检测[1]。误用检测技术需要事先建立入侵行为特征库, 入侵检测时采用特征匹配的方法确定是否存在入侵行为。异常检测技术需要事先通过训练数据集建立正常行为模型, 入侵检测时根据是否显著偏离正常模型判断是否存在入侵行为。因此, 异常检测可以检测出未知的入侵行为, 在实际中应用较多[2]。

通讯作者: 白琳, 硕士, 副教授, Email: bailin@xupt.edu.cn。
本课题得到西安市科技创新引导项目（No.201805040YD18CG24(7)）资助。
收稿日期: 2019-05-21; 修改日期: 2019-08-22; 定稿日期: 2019-08-23

近十几年来, 在许多异常检测系统中, 聚类分析方法以其无监督性被广泛采用。各种聚类算法及其改进被用于训练数据集, 建立正常模型[3-4]。在日益复杂的网络大数据环境下, 对聚类算法的要求也越来越高: 能够处理海量、异构、混合属性的数据, 与数据分布无关, 自适应性、自学习性、拓展性能好。聚类和智能技术的融合为提高聚类算法的性能提供了良好的基础和思路。

2 智能聚类技术在入侵检测中的应用

聚类分析是将未做标记的样本集按照特定的准则划分成若干个子集, 并保证相似的样本尽量在同一个子集中。相似度的度量依靠数据对象描述属性的取值来确定[5]。

传统聚类算法, 如: K-MEDOIDS、 K-MEANS、EM、BIRCH 一般只对小样本有效, 可拓展性差, 并且对初始化敏感、依赖聚类原型, 容易陷入局部最优。为此, 有研究者将传统聚类算法和智能计算方法融合, 设计了一系列智能聚类算法。如: 基于遗传算法的聚类、免疫进化聚类[6]、克隆选择聚类[7]以及进化人工免疫网络[8]和基于神经网络的免疫网络[9]。这些方法在提高聚类算法的可拓展性、自适应性等方面效果显著。但也存在较为严重的缺陷: 遗传聚类容易早熟; 免疫聚类算法中, 疫苗的提取依赖先验知识, 实际操作中难度较大; 克隆聚类中的克隆规模依赖经验值, 无法自适应调整; 进化免疫网络和基于神经网络的免疫网络无法有效处理边界数据和噪声数据。

这些智能聚类方法中所采用的学习策略, 其仿生机制都是源于神经系统或免疫系统。但在实际中, 生物体内在拥有三大系统, 包括神经、免疫和内分泌系统。这三大系统相互作用与反馈、协调与制约, 促进与拮抗, 形成一个有机的整体共同维持着机体的内稳态。机体内的这种调控功能, 就来源于内分泌系统的激素调节机制。即内分泌激素可以调控机体的神经和免疫系统功能。

因此, 将人工内分泌计算中的激素调节机制引入免疫系统计算方法, 用于调节和刺激抗体群的进化和免疫学习的过程, 使抗体进化过程中的克隆规模可以自适应的、动态的自调节, 有效提高聚类精度和性能。

3 人工内分泌计算

3.1 内分泌系统

内分泌系统是生物体的重要控制系统, 主要通过激素对机体进行调节, 激素由内分泌腺体分泌, 释放到血液中, 随血液循环到达生物体全身, 对整个机体起到调节作用[10]。

人工内分泌系统(Artificial Endocrine System, AES), 是一种借鉴生物体内分泌系统的信息处理机制, 将其应用于计算、控制、通信等领域而形成的模型或系统[11]。

机体内, 内分泌系统和免疫系统间的相互作用与协调, 主要体现在: 激素通过免疫细胞受体使其免疫功能增强或减弱; 而免疫系统通过细胞因子对内分泌系统发生作用[9]。即: 内分泌系统调节机体的免疫功能, 免疫系统再对其进行应答和反馈。

因此, 将内分泌计算和免疫计算相融合, 采用内分泌激素调节算子对进化免疫网络的抗体克隆规模和网络压缩规模进行调节, 实现网络进化学习的自适应、动态调节, 加速抗体成熟过程和网络收敛过程。

3.2 内分泌激素调节算法

L. S. Farhy 于 2001 年提出了激素调节的 Hill 函数[12]。即激素分泌的上升调节函数和下降调节函数。

$$F_{\text{up}}(G)=\frac{G^n}{T^n+G^n} \tag{1}$$

$$F_{\text{down}}(G)=\frac{T^n}{T^n+G^n} \tag{2}$$

其中，F 是 Hill 调节函数, up 表示刺激、down 表示抑制; G 是自变量, $n(n\geqslant 1)$ 是 Hill 系数, $T(T>0)$ 是阈值参数, n 和 T 代表曲线上升或是下降的斜率, 对应激素刺激或抑制的改变速度。

且, Hill 函数满足:

$$\begin{cases} F_{\text{up}}(G)=1-F_{\text{down}} \\ F(G)_{G-T}=\dfrac{1}{2} \\ 0\leqslant F(G)\leqslant 1 \end{cases}$$

4 人工免疫系统

人工免疫系统(Artificial Immune System, AIS), 是模拟生物体免疫系统工作过程的一种信息处理机制。主要算法包括克隆选择算法(Clonal Selection Algorithm, CSA) 和人工免疫网络(Artificial Immune Network, AiNet)。

4.1 克隆选择算法

克隆选择算法的中心思想是: 在抗原的刺激下, 机体的免疫细胞会通过克隆操作而增殖, 再通过遗传、交叉和变异操作分化为多样性效应细胞和记忆

细胞。在群体控制策略下, 与抗原间亲合度低的抗体在经过进化算子操作后, 亲合度会趋向于成熟。

$$A(k)\xrightarrow{\text{clone}}A'(k)\xrightarrow{\text{immune genic operation}}A''(k)\xrightarrow{\text{selection}}A(k+1) \tag{3}$$

克隆算法包括克隆、免疫基因操作、克隆选择三个步骤, 见公式(3)。克隆的实质是在一代进化过程中, 根据亲合度的大小, 在侯选解的附近产生一个变异解群体, 以此扩大搜索范围、增加抗体群多样性, 防止早熟和搜索陷于局部极小值[13]。

4.2 人工免疫网络

AiNet是一种用于聚类分析的免疫系统方法, 其主要思想是: 设$X=\{x_1,x_2,\cdots,x_n\}$是聚类样本集。$x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^t$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, S作为抗原, 可决定抗体—抗体、抗原—抗体间的相互作用[14]。这种相互作用可用一个连通图表示。那么, 免疫网络就是一个加权的图 G, 由一组不完全连接的抗体节点组成。

AiNet 做为一种网状结构的聚类分析方法, 聚类前无需指定类别数, 与数据分布无关, 不依赖聚类原型, 是真正意义上的无监督算法。但是, 如果数据子集的边界比较模糊, 或者样本集本身就存在噪声, 这种特殊抗原会极强的刺激免疫反应, 引起细胞增殖, 导致进化网络出现结构不清晰的问题。

如果抗体网络的免疫反应能够自适应、动态的进行调节, 就能有效控制特殊抗原对网络结构的干扰, 推进网络进化学习过程的自适应性和自学习性, 从而有效提高聚类性能和收敛速度。

5 基于激素调节的免疫网络

5.1 算法思想

将人工内分泌系统的激素调节机制引入免疫网络的抗体种群进化过程, 利用亲合度函数动态调节抗体的克隆规模和网络压缩规模, 充分发挥优秀个体的先进特性来刺激亲合度成熟, 动态调控抗体网络的进化、学习过程。同时, 为适应激素的调节机制, 免疫基因操作也要做相应改进, 为体现局部搜索在抗体群进化过程中的自适应能力, 本文采用一致变异算子。

5.2 相异度函数

本文 benwen 实验采用的 KDD 入侵检测训练数据, 包含数值型和类属型, 是具有混合属性的网络连接记录。本文采用下列相异度测量函数来描述混合属性:

$$d(x_i,x_j)=\sqrt{\sum_{k=c_1}^{c_{m_c}}(x_{ik}-x_{jk})^2+\lambda\sum_{l=d_1}^{d_{m_d}}\delta(x_{il},x_{jl})} \tag{4}$$

常数λ用来调节两种属性在目标函数中的比例。$\delta(\cdot)$定义为: $\delta(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases}$。

5.3 相关定义

聚类目标函数:

本位的聚类目标函数采用 $C(\boldsymbol{W},\boldsymbol{P})$。在此, 需要求解样本集 X 的划分矩阵 $\boldsymbol{W}$ 和原型矩阵 $\boldsymbol{P}$。由于 $\boldsymbol{W}$ 和 $\boldsymbol{P}$ 相关, 本文求出 $\boldsymbol{P}$ 即可。

抗体: 一组聚类原型 $\boldsymbol{P}$。

抗体网络: 矩阵 $\boldsymbol{A}(N_A\times p)$。其中, $\boldsymbol{A}$ 的每个行向量代表一个抗体节点。

抗体—抗原亲合度函数:

根据目标函数越小, 聚类效果越好, 合度越大的原则, 构造抗原$x_i(i=1,\cdots,n)$与抗体$y_j(j=1,\cdots,m)$亲合度函数:

$$f(x_i,y_j)=\frac{N}{1+\sum_j\left\|y_j-x_i\right\|+\lambda\times\log_a(a-1+N)\times\sum_j\delta(y_j,x_i)} \tag{5}$$

抗体—抗体亲合力:

亲合力s_{ij}为抗体间距离测度, $s_{ij}=d(y_i,y_j)$。s_{ij}越小, y_i与y_j差异越小。

5.4 基于激素调节的克隆算子

5.4.1 克隆算子

抗体的克隆操作为:

$$T_c^C\left(y_{rm}\right)=\boldsymbol{I}_m\times y_{rm},\quad m=1,2,\cdots,k \tag{6}$$

其中, $\boldsymbol{I}_m$是元素为 1 的 q_m 维行向量。

$$q_m=Int(n_c*\frac{f(x_i,y_{rm})}{\sum_{b=1}^{k}f(x_i,y_{rm})}),\ m=1,\cdots,k \tag{7}$$

$Int(c)$表示大于 c 的最小整数, n_c 为克隆总规模。

单个抗体的 q_m 大小取决于 f。因此, 单个抗体的克隆规模完全由亲合度决定。

5.4.2 激素调节克隆算子

抗体克隆的激素调节操作为:

对于第 t–1 代种群, 其平均亲合度表示为:

$$\overline{f}_{t-1}=\sum_{i=1}^{N}f_{t-1}^i/N \tag{8}$$

f_t^i 代表第 t 代中的第 i 个个体亲合度。

种群的多样性函数为:

$$D(t)=\sum_{i=1}^{N}(f_t^i-\bar{f}_{t-1})^2/N \tag{9}$$

下降规律激素调节函数为:

$$F(t)=\frac{N}{N+\sum_{i-1}^{N}(f_t^i-\bar{f}_{t-1})^2} \tag{10}$$

令:

$$G=\sum_{i-1}^{N}(f_t^i-\bar{f}_{t-1})^2 \tag{11}$$

上述公式描述了当代抗体种群中, 个体亲合度与上一代平均亲合度之间的关系。

当代抗体群中, 由函数 $D(t)$和 $F(t)$共同调节个体克隆规模, 第 t 代个体的克隆规模 $q(t)$为:

$$q_i(t)=\mathrm{int}(\eta\times D(t)\times F(t))=Int(\eta\times\frac{\sum_{i=1}^{N}(f_t^i-\bar{f}_{t-1})^2}{N+\sum_{i=1}^{N}(f_t^i-\bar{f}_{t-1})^2}),\quad i=1,2,\cdots,N \tag{12}$$

参数 η 依据经验值而定。

激素调节克隆后的种群变为:

$$A^{'}(t)=\{A(t),A_1^{'}(t),A_2^{'}(t),\cdots,A_N^{'}(t)\} \tag{13}$$

其中,

$$A_i^{'}(t)=\{A_{i1}(t),A_{i2}(t),\cdots,A_{iq(t)-1}(t)\},j=1,\cdots,q_i(t)-1 \tag{14}$$

5.4.3 非一致变异

为了最大限度的保留最佳个体, 改进较差个体, 对激素调节克隆后的抗体按变异概率实施非一致变异操作, 以此来体现局部搜索求解的自适应能力。

根据概率 P_m 产生父代, 将个体编码的各个基因作为变异点, 对其进行如下操作: 设某一个父代个体为 $y_i^t=(y_1^t,y_2^t,\cdots,y_k^t,\cdots,y_{n\times l}^t)$, 由 y_i^t 向 $y_i^{t+1}=(y_1^t,y_2^t,\cdots,y_k^{t+1},\cdots,y_{n\times l}^t)$ 变异时, 若变异点的参数 y_k^t 的变化范围为 $\left[U_{\min}^k,U_{\max}^k\right]$, 则新基因值 y_k^{t+1} 由(15)式确定:

$$y_k^{t+1}=\begin{cases}y_k^t+\Delta(t,U_{\max}^k-v_k),若random(0,1)=0;\\ y_k^t-\Delta(t,v_k-U_{\min}^k),若random(0,1)=1.\end{cases} \tag{15}$$

$random(0,1)$表示按等概率取 0 或 1; v_k 为微小扰动量; $\Delta(t,z)=z\times(1-r^{(1-t/T)^b})$ 表示在[0,z]区间非均匀分布的随机数, r 为[0,1]内的随机数, b 为系统参数, 本文取 2, T 为最大进化代数。

5.5 激素调节免疫网络学习算法

算法流程框图见图 1。

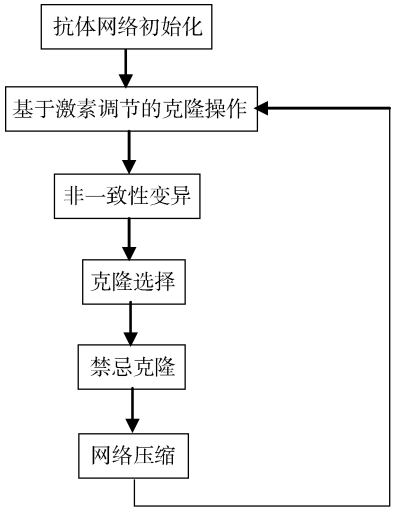

图 1 激素调节免疫网络流程框图

Figure 1 Flow chart of hormone-regulated AiNet

算法描述:

Step1 初始化: 随机生成初始网络, 用矩阵 $\boldsymbol{A}(N_A\times p)$ 表示抗体网络, 网络节点即抗体 $y_j(j=1,\cdots,N_A)$, 设置各参数和迭代终止条件;

Step 2 对每个输入的抗原 $x_i(i=1,2,\cdots,N)$, 进行下列操作:

Step2.1 根据公式(4)计算抗体-抗原亲合度;

Step2.2 激素调节克隆: 选择 k 个亲合度最高的抗体($y_{r_1},y_{r_2},\cdots,y_{rk}$), 根据 § 5.4.2 的操作进行抗体的激素调节克隆操作 $q_i(t)$;

Step2.3 非一致变异: 对激素调节克隆后的抗体按 § 5.4.4 进行变异操作;

Step2.4 计算抗体-抗原亲合度, 选 λ%亲合度最高抗体组成记忆单元, 得到矩阵 $\boldsymbol{A}_P$。

Step2.5 删除 $\boldsymbol{A}_P$ 中亲合度小于门限 б 的抗体节点;

Step2.6 按§5.3 计算 $\boldsymbol{A}_P$ 中抗体-抗体亲合力, 删除其中亲合力小于门限 μ 的节点;

Step2.7 将 $\boldsymbol{A}_P$ 加入 $\boldsymbol{A}$ 中;

Step 3 禁忌克隆: 计算 $\boldsymbol{A}$ 中每个抗体 y^*与其他抗体的亲合力, 对任一抗体, 计算与该抗体亲合力小于门限 μ 的抗体个数 e, 若 e 小于门限 σ_f, 删除该抗体;

Step 4 网络压缩: 计算 $\boldsymbol{A}$ 中抗体-抗体亲合力,

删除其亲合力小于门限 μ 的结点, 按比例 r%随机选择抗体代替已删除抗体加入网络 $\boldsymbol{A}$ 中, 得到新网络;

Step 5 若满足终止条件(当前最优抗体连续 10 代无改进), 转 Step 6, 否则对抗体群中比例为 β2 的具有最小亲合度的抗体, 重新初始化, 返回 Step 2。

Step 6 网络输出: 算法停止。

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结点间的相互联系并描述网络结构的内部亲合矩阵组成。为了获得数据样本集内在的聚类结构, 对上述输出的抗体网络, 采用最小生成树的方法进行聚类分析。

5.6 聚类分析

通过网络进化学习得到的网络$\boldsymbol{A}$是聚类样本X的空间特征映射, 对聚类样本的聚类分析即转换为对记忆网络的聚类分析, 对记忆网络采用简化的连通图最小生成树来实现聚类分析[15]。

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, 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 对抗体网络$\boldsymbol{A}$构建最小生成树MST;

Step 2 在上述MST上, 进行剪枝操作, 将权值大于剪枝系数σ_m的边剪断;

Step 3 剪枝后, MST由多棵相互断开的子树组成, 每棵子树的抗体节点就是一个聚类子集;

Step 4 用 Bar 图来描述 MST 的边长, Bar 图中山谷的数目就是最终的聚类类别数。

6 基于激素调节免疫网络的异常检测

6.1 正常模型

对训练数据集进行聚类分析后, 抗体群被划分为 h 个子集 $P_c(1\leqslant c\leqslant h)$。每个子集内部, 抗体间的距离较近, 不同子集的抗体间距离较远。抗体子集的分类情况映射了训练集在样本空间中的分布。

对训练集$X=\{x_1, x_2, ..., x_n\}$的分类, 就是为每个抗体子集贴上标签。计算 x_i $(1\leqslant i\leqslant n)$和不同抗体子集的距离, 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。

对分类后的数据集确定正常类或异常类。由于本文的检测系统基于两个合理的假设[5]:

(1) 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;

(2) 入侵行为比正常行为本质特性差异很大且相对很少。

因此, 根据获得的样本分布就可以划分各正常类和异常类。若某个类的数据量与样本总数据量之比不小于$\eta(0<\eta<1)$, 为正常类。正常类中的抗体节点就是正常数据的代表点(正常模型)。

6.2 检测算法

Step 1 对数据集 $Y=\{y_1, ..., y_n\}$, 计算 $y_i(1\leqslant i\leqslant n)$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y_i的代表点,根据p_m所属模型的类别标签判断 y_i 是否为异常数据。

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

6.3 仿真实验

6.3.1 实验数据

仿真实验数据选自 KDD CUP[16]数据集, 它包含了 9 个星期的网络流量信息。其中, 7 周的训练数据集包含约 500 万条网络连接记录, 2 周时间的测试数据集约包含约 200 万条连接记录。连接记录有 42 个属性(其中有 33 个连续属性和 8 个离散属性), 具体如持续时间、协议类型、传输的字节数、标签等。

标签属性标识了该条网络连接记录是正常的或是某种具体的攻击。训练数据有 22 种攻击类型, 测试数据有 37 种攻击类型, 因此 15 种未出现在训练集的攻击可作为训练集的未知攻击。攻击类型被分为 4 大类: 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)[17]。

实验训练和测试数据集构造见表 1 和表 2。

表 1 实验数据集

Table 1 Experimental data set

数据集	数据量(条)	攻击数据量(条)	攻击类型(种)	未知攻击(种)
训练集	180,000	2,000	22	\
测试集 1	100,000	1,000	33	11
测试集 2	100,000	1,100	37	15

表 2 数据集各种攻击类型数

Table 2 Number of attack types in dataset

数据集	DOS(条)	U2R(条)	R2L(条)	Probing(条)
训练集	880	270	460	390
测试集 1	190	230	260	320
测试集 2	450	110	360	80

6.3.2 数据预处理

1. 连续属性离散化

本文采用等分区间法将数据连续属性的取值区

间[a, b]划分为 N 个小区间, 每个小区间对应一个离散值:

$$(a, a+(b-a)/N), \cdots, (a+(b-a)/N' a+2(b-a)/N), \cdots, \left(b-(b-a)/N'^{b}\right) \tag{16}$$

2. 枚举类型

本文采用整数表示字符枚举类型属性。如: 0 表示“http”协议, 1 表示“ftp”协议。

6.3.3 仿真实验

使用本文的激素调节免疫网络对训练数据集进行聚类分析, 再利用最小生成树映射出聚类结果。见图 2, 图中 Bar 图中山谷的数目就是聚类类别数 5。

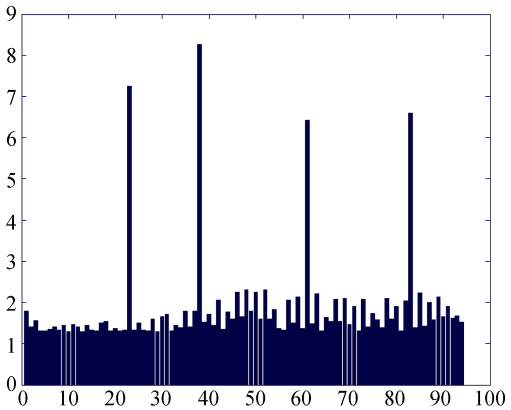

图 2 本文算法的聚类 Bar 图

Figure 2 Clustering Bar graph

按 § 6.1 的贴标签方法为 5 个子类贴上标签。

根据入侵检测系统的两个假设, 按各个子类的数据量确定其为正常或异常: 若某个子类的数据量与样本数据总量之比不小于 $\omega(0<\omega<1)$, 为“正常”, 否则为“异常”。

对测试集 X, 计算每个 x_i 到每个子类的距离, 找出最短距离 d_{min}。最短距离对应的子类若为异常, 则将 x_i 判为入侵。如果 $d_{min}\geqslant\alpha$, x_i 就是未知类型的入侵。由此, 实现了异常检测。

6.3.4 实验结果及分析

将传统 C-均值聚类算法(算法 1)、基于 CSA 的聚类算法[5](算法 2), AiNet 算法(算法 3)作为本文算法的对比方法。激素调节参数 η=2.5, 变异概率 p_m=0.25, 选择最佳个体百分比设为 18%, σs =0.25。

30 次独立实验取平均值, 本文算法和算法 2 得到聚类结果为 5 类, 算法 1 为 6 类。三种算法的标类结果见表 3。

表 3 标类结果

Table 3 Category label results

类别 / 算法	1	2	3	4	5	6
本文算法	正常	异常	异常	异常	异常	\
算法 1	正常	异常	异常	异常	异常	\
算法 2	正常	正常	异常	异常	异常	异常
算法 3	正常	异常	异常	异常	异常	\

入侵检测结果见表4, (已知入侵指训练样本中包含的22种攻击; 未知入侵指训练集中未包含的15种攻击)。其中, 检测率指被检测出来的异常数据占异常数据总数的百分比。误警率指正常的数据被错误的判断为异常的数目占正常数据总数的百分比。

具体的, 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; 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 检测结果

Table 4 Detection result

数据集	算法	检测率(%)		误警率(%)
		已知攻击	未知攻击	
测试集 1	本文算法	92.35	85.22	3.89
	算法 1	68.15	55.77	8.68
	算法 2	85.25	78.22	4.91
	算法 3	80.17	65.33	7.35
测试集 2	本文算法	91.33	87.27	3.61
	算法 1	66.23	52.39	8.15
	算法 2	84.64	80.11	5.23
	算法 3	81.12	70.23	7.01

表 5 给出了本文算法对 4 大类攻击的检测结果。

表 5 各种攻击检测结果

Table 5 Detection results of various attacks

数据集	测试集 1 检测率(%)		测试集 2 检测率(%)	
攻击类型	已知攻击	未知攻击	已知攻击	未知攻击
DOS	89.33	78.79	90.01	85.64
U2R	97.25	85.36	98.37	90.33
R2L	87.45	83.52	80.25	88.90
Probing	95.37	93.21	96.69	84.21

从表 4 可以看出, 本文算法的检测效果更加理想, 优势比较明显, 并且能够有效检测出未知类型的攻击。说明激素调节下的免疫网络, 具有更优的聚

类性能。具体分析如下:

(1) 传统方法(C-均值)对初始化敏感、依赖聚类原型、算法可拓展性和鲁棒性较差, 不能有效处理大规模数据, 所以其检测效果较基于生物计算的智能算法有明显差距。

(2) 借鉴生物体内稳态平衡机制, 将生物进化策略与聚类分析过程融合, 利用内分泌的调节机制调控进化算子和免疫基因操作的过程, 使得激素调节算子在抗体的克隆过程中发挥了重要的调节作用, 所以每代抗体的克隆增殖能够自适应的调整到最佳规模, 不仅有利于进化过程中的种群多样性, 并且能够最大化的发挥优秀个体的性能, 抑制不良个体对进化学习的影响。刺激和加速抗体亲合度成熟的过程。同时, 激素调节也作用于网络规模的动态调整, 保证了网络的进化是动态的、自适应的和自学习的, 从而最真实的映射出样本空间的分布。最终保证算法整体快速、高效的收敛至全局最优。

(3) 本文算法的非一致变异算子和禁忌克隆算子, 有助于提高网络的学习性能。使其动态性能和结构受进化策略和免疫学习策略控制。非一致变异体现了搜索的局部求解自适应能力, 保留最佳个体、改进较差个体。在激素调节克隆算子的共同作用下, 保证了本文算法在能够增加解的多样性的基础上, 将全局搜索和局部搜索有机结合, 推进全局寻优。

(4) 距离尺度分析下的网络连接记录样本集, 其数据集边界会存在模糊不清的现象, 或出现噪声点。克隆算子、禁忌克隆算子和网络压缩操作经过激素算子的动态调节, 网络的免疫耐受性和免疫特异性有所提高, 因而能够有效克服边界模糊或噪声对聚类效果的影响。

(5) 从表 4 和 5 可以看出, 本文算法对未知攻击类型的检测率不是很高, 说明网络学习的自适应性和自调节能力以及距离尺度的动态调整方面还存在一定的局限性, 有待进一步研究和改进。

7 结论

生物体是一个复杂的、协调的系统。其中包含的内分泌、免疫、神经等相关生命机理为智能聚类技术的发展提供了良好的借鉴。本文将人工内分泌系统的激素调节机制引入改进的免疫网络中, 通过激素调节克隆算子和网络规模。保证在群体的进化过程中, 充分发挥优秀个体的积极作用, 抑制不良个体的消极作用, 刺激亲合度趋向于成熟。增强了聚类算法的自适应性、自学习性和自稳定性, 使其具有更高的全局寻优特性和更快的收敛速度。在此基础上, 通过本文的新聚类算法对训练数据集进行聚类分析得到了入侵检测的正常模型, 实现了测试数据的异常检测。检测效果理想, 并能检测出未知攻击。本文的异常检测系统是可行的、有效的, 具有一定的使用价值。

参考文献

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

[2] 李威, 杨忠明. 入侵检测系统的研究综述[J]. *吉林大学学报(信息科学版)*, 2016, 34(05): 657-662.

[3] 江颉, 王卓芳, 陈铁明, 朱陈晨, 陈波. 自适应 AP 聚类算法及其在入侵检测中的应用[J]. *通信学报*, 2015, 36(11): 118-126.

[4] 肖敏, 韩继军, 肖德宝, 吴峥, 徐慧. 基于聚类的入侵检测研究综述[J]. *计算机应用*, 2008(S1): 34-38+42.

[5] 崔文科. 基于聚类算法的入侵检测系统的设计与实现[D]. 电子科技大学, 2016.

[6] 刘静, 钟伟才, 刘芳等. 免疫进化聚类算法[J]. *电子学报*. 2001, 29(12): 1869-1872.

[7] 李洁, 高新波, 焦李成. 一种基于 CSA 的混合属性特征大数据集聚类算法[J]. *电子学报*, 2004, 32(3): 367-372.

[8] Leandro Nunes de Castro, Fenando J. Von Zuben. An Evolutionary Immune Network for Data Clustering. *Proc. of the IEEE SBRN*, 2000, 84-89.

[9] 白琳. 基于神经计算和进化网络的入侵检测[D]. 西安电子科技大学, 2005.

[10] 林广栋, 王煦法. 人工内分泌系统调节人工神经网络的控制模型[J]. *中国科学技术大学学报*, 2012, 42(02): 148-153+160.

[11] 林广栋. 人工内分泌系统新机制及应用研究[D]. 合肥: 中国科学技术大学, 2012.

[12] 王祎, 陈为栋, 顾幸生等. 基于内分泌激素调节机制的免疫算法的 Flowshop 调度问题[J]. *系统仿真学报*, 2008, 20(13): 3425-3430.

[13] L S Farhy. Modeling of oscillations of endocrine networks with feedback [J]. *Methosd Enzymol (S0076-6879)*, 2004, 384: 54-81.

[14] Haifeng DU, Licheng JIAO, Sun' an Wang. Clonal operator and anti2 body clonal algorithm [A]. Proceedings of the First International Con2 ference on Machine Learning and Cybernetics [C]. USA: IEEE Press, 2002. 506-510.

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

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

[17] 张新有, 曾华燊, 贾磊. 入侵检测数据集 KDDCUP99 研究[J]. *计算机工程与设计*, 2010, 31(22): 4809-4812+4816.

白　琳　于2005年在西安电子科技大学计算机应用专业获得硕士学位。现任西安邮电大学计算机学院副教授。研究领域为智能信息处理。研究兴趣包括：智能数据挖掘、自然语言处理等。Email: bailin@xupt.edu.cn

杨　超　于 2008 年在西安电子科技大学大学密码学专业获得工学博士学位。现任西安电子科技大学，网络与信息安全学院教授、博士生导师。研究领域为网络与信息安全。研究兴趣包括：网络流量隐私分析与攻击检查、物联网协议分析与设计等。Email: chaoyang@xidian.edu.cn