文章编号: 1001-0920 (2011) 10-1562-05

邻域形态空间与检测算法

张凤斌, 席 亮, 王大伟, 岳 新

(哈尔滨理工大学 计算机科学与技术学院，哈尔滨 150080)

摘 要: 针对免疫入侵检测中实值空间存在的问题, 借助免疫细胞的表位组织形式和离散拓扑理论, 提出一种新的形态空间表示法——邻域表示法. 该方法利用数据的集合特性, 采用空间中互不相交的邻域表示自体/检测器, 并设计匹配策略和检测算法. 实验表明, 邻域空间可以较好地弥补实值空间的缺陷, 提高检测器生成效率, 改善系统整体检测效果.

关键词: 免疫原理；入侵检测；实值；邻域

中图分类号: TP273 **文献标识码:** A

Neighborhood shape-space and detection algorithm

ZHANG Feng-bin, *XI Liang*, *WANG Da-wei*, *YUE Xin*

(College of Computer Science and Technology，Harbin University of Science and Technology，Harbin 150080, China．Correspondent：XI Liang，E-mail：xljyp2002@yahoo.com.cn)

Abstract: For the problem of shape-space in the immunity-based intrusion detection system(IIDS), by using the epitope constitute of immune cells and discrete topology theory, a new method, named neighborhood is proposed. Furthermore, the corresponding matching rules and detectors generating algorithm are given. Experimental results show that, the new shape-space can solve the defects of real-valued representation, increase the efficiency of detector generation effectively, and improve the system's detection rate.

Key words: immunity principle；intrusion detection system；real-valued；neighborhood

1 引 言

入侵检测技术作为异常检测的一种, 是当今计算机安全领域一项重要研究课题, 而生物免疫机制与异常检测一样, 都是将观察物通过一定策略区分为自体(正常)与非自体(异常), 且生物免疫中的自适应控制、鲁棒控制等特性可以很好地作用于入侵检测系统, 为入侵检测技术的研究提供了一个新思路[1-3]. 自Forrest[4]将生物免疫机制引入到计算机安全以来, 基于免疫的入侵检测研究取得了巨大成绩. 如否定选择算法的创立、二进制形态空间和实值形态空间[5]、克隆选择算法[6-7]和树突状细胞算法的应用[8]等. 但是由于忽视了对于形态空间合理性的论证以及自体集分析, 目前该领域的研究遇到了许多棘手问题.

针对实值空间存在的问题, 深入理解免疫细胞的表位组织形式与官能, 借鉴离散拓扑理论, 提出一种新的形态空间表示法——邻域形态空间. 设计相应的匹配规则、检测算法等, 拟达到克服实值表示局限性, 特别是实值空间维数灾难问题和解放单维属性独立性的目的. 从而使系统更好地对网络行为特征提取和判定, 提高系统的整体性能.

2 实值形态空间问题分析

实值表示法将与R^n子集对应的"自体/非自体"形态空间, 归一化到$[0,1]^n$超矩形空间或直径为1的超球体空间U内. U分为自体子空间S和非自体子空间R, 即$U = S \bigcup R$. 检测器集$D \subseteq R$, 自体s和检测器d分别分布于S和D, 一般通过否定选择算法进行区分. 自体/检测器认为是n维超球体: 一个n维空间的点及其半径(判定阈值).

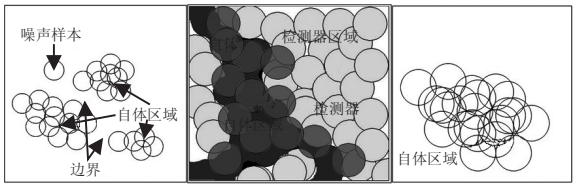
1) S多分区与噪声样本问题. 通常认为S是一个整体区域, 然而不同范畴内的样本属性值分布于不同的值域内, 即使在同一属性域, 其值也可能分布于不同值域内, 即自体子空间是多分区的. 并且由于自体收集时不可避免地采集到噪声样本, 根据实值空间的

收稿日期: 2010-06-24；**修回日期:** 2010-08-26.

基金项目: 国家自然科学基金项目(60671049, 61172168).

作者简介: 张凤斌(1965−), 男, 教授, 博士生导师, 从事网络与信息安全等研究；席亮(1983−), 男, 博士生, 从事网络与信息安全的研究.

特点, 噪声多位于分区边缘区域. 图1(a)是二维实值自体多分区问题和噪声样本现象. 在训练检测器阶段, 噪声样本占据的非自体空间不能被检测器覆盖, 造成检测器黑洞.



(a) S多分区和噪声样本　(b) S/R边界交叉　(c) 自体高重叠率

图 1　实值空间问题

2) 边界交叉问题. 实值空间采用单点表示自体/检测器, 以检测阈值(半径)来判定异常行为. 由于半径的存在, 位于自体边界的自体识别区域会超过自体边界而覆盖部分非自体区域, 位于非自体边界的检测器识别区域也会超出非自体边界而覆盖部分自体区域. 这种现象称为S/R边界交叉, 如图1(b)所示. 在检测阶段不仅会降低检测率, 而且会造成系统大量的漏报.

3) 自体样本高重叠率问题. 在某些自体分区内样本间差别很小, 由于实值空间的特点, 其位置非常近, 识别区域重叠覆盖率较高, 如图1(c)所示. 另外, 通过推导可得候选检测器d_0匹配S失败的概率(经过耐受的概率)为$p_{tf}=(1-p_m)^{N_s}$, 检测阶段一次漏报的概率为$p_{tf}=(1-p_m)^{N_d}$. 其中: p_m为d_0与任一s的匹配概率, N_s为自体数, N_d为检测器数. 当N_d足够大时, $p_{df}=\mathrm{e}^{-N_d p_m}$, 从而可得$N_s$与$d_0$数量$N_{d_0}$间的关系为

$$N_{d_0}=N_d/p_{tf}\approx\frac{-\ln(p_{df})}{p_m(1-p_m)^{N_s}}. \tag{1}$$

由式(1)可知, 当p_{df}和p_m一定时, N_{d_0}和N_s为指数级关系: 在训练阶段, 自体越多, 需要的候选检测器越多, 系统训练检测器的代价越大. 自体高重叠率现象说明S中存在较多多余自体, 这势必影响训练检测器的效率.

4) 检测器维数灾难问题. 目前实值空间主要匹配策略根据样本点间的Euclidean距离或Manhattan距离来判定样本间的相似度. 而Euclidean距离和Manhattan距离分别是Minkowski距离参数q为2和1时的特例. 以Euclidean距离为例, 给出实值空间里半径为r的n维超球体的体积为

$$V(n,r)=r^n\frac{\pi^{n/2}}{\Gamma(n/2+1)}, \tag{2}$$

其中

$$\Gamma(n/2+1)\approx\sqrt{2\pi}\mathrm{e}^{-n}n^{n+1/2}. \tag{3}$$

从而有

$$\lim_{n\to\infty}V(n,r)=\frac{1}{\sqrt{2\pi}}\lim_{n\to\infty}\frac{(r\mathrm{e}\sqrt{\pi})^n}{n^{n+1/2}}=0. \tag{4}$$

由式(4)可知, 对于r固定的超球体, 当$n\to\infty$时, 体积趋于0. 即对于实值检测器, 随着属性数的增加, 其体积(识别区域)趋于0, 限制了实值检测器属性域的选取. 对于复杂的网络环境, 要求检测器以低维属性域来监测显然无法满足实际需要.

设形态空间是半径为r的超球体, 则半径在$[r-\varepsilon,r]$范围内的超球环的体积占整个形态空间体积的比例为

$$\lambda=1-\frac{V(n,r-\varepsilon)}{V(n,r)}=1-(1-\varepsilon/r)^n, \tag{5}$$

则有

$$\lim_{\varepsilon\to r}\lambda=1-\lim_{\varepsilon\to r}(1-\varepsilon/r)^n=1. \tag{6}$$

由式(6)可知, 大部分检测器将分布在超球体半径接近1的超球环内, 而内部区域将很少被覆盖以致造成许多空洞, 导致实值空间以Minkowski体系距离为衡量标准的实际可操纵性大打折扣.

3　邻域形态空间及相关算法

在生物免疫细胞中, 表位组织形式的多样性决定了其识别抗原的多样性. 每个免疫细胞可以看作是多个表位的集合, 特定表位组织形式针对特定抗原的识别性决定了免疫细胞的特异性和多样性. 实值表示法忽略了自体/检测器各属性的特殊性, 也忽略了自体/检测器的集合性, 只是简单通过Minkowski距离体系下的坐标系将各样本集合在一起, 通过半径加以区分, 从而造成了以上问题. 为此, 根据数据的集合性以及离散拓扑理论, 借鉴免疫细胞表位组织形式与官能性, 从表位思想提出一种新型形态空间——邻域形态空间: 以空间中互不相交的邻域来表示自体/检测器, 并设计相应的检测算法来训练检测器并检测异常.

3.1　邻域形态空间和匹配规则定义

定义 1　集合X上的一个拓扑是X的子集的一个族T, 它需满足: $\varnothing$和X在T中; T中任意子族的元素的并和交在T中. 一个制定了拓扑T的集合X称为一个拓扑空间(T,X). 如果X为任意一个集合, 其所有子集的族是X的一个拓扑, 则称为离散拓扑.

定义 2　X是带有拓扑T的一个拓扑空间, 如果子集U是族T的一个元素, 则U是X中的一个开集. 如果U是包含元素x的一个开集, 则称U为x的一个邻域, 记作$U(x)$.

直观的说, 空间中点x的邻域$U(x)$是包含该点的集合, 该点周围的点集合包含于$U(x)$, 这与自体/检测器的集合性很相似. 以归一化后的实值向量表示一个点, 则同一类型的点分布于空间的一个区域组

成一个点集合, 其邻域的并组成了该点集合的一个邻域, 各点集合的邻域的并组成邻域形态空间, 则每个邻域都是集合 $[0,1]^n$ 内的一个超立方体. 由于 U 是包含元素 x 的一个开集, 超立方体不是其任何一个角的邻域, 对于一个邻域而言, 每一维都是一个开集 (a,b), 其中 $0<a<b<1$.

邻域表示法将集合 $[0,1]^n$ 分解成多个互不相交的邻域. 对于集合的第 i 维 $[0,1]_i$, 将其分为 m_i 个子集, 彼此互不相交, 即 $\bigcup_{j=1}^{m_i}(x_{i,j-1},x_{i,j})$, 其中 $x_{i,0}=0$, $x_{i,m_i}=1$. 扩展到 n 维空间, 则形态空间就是所有邻域的集族, 可以表示为 $\bigcup(x_{j-1},x_j)$, 其中 $1\leqslant j\leqslant m_i$, $i=1,2,\cdots,n$, 每个 (x_{j-1},x_j) 即为自体/非自体(检测器). 综上所述, 邻域形态空间是集合 $[0,1]^n$ 的一个子集族, 其中每个子集都是一个开集, 并且是该子集包含元素的一个邻域, 具有以下性质:

1) 空间内任意子集的交为 $\varnothing$;

2) 每一个元素都是集合 $[0,1]^n$ 内的一个超立方体, 且不包含任意顶点;

3) 空间内全部元素和顶点的并等于整个集合 $[0,1]^n$.

邻域空间的自体/检测器与其他空间下的自体/检测器一样, 其属性值需要正规化处理. 不同于实值空间的处理办法, 领域空间针对不同类型的属性将采用不同的正规化方法:

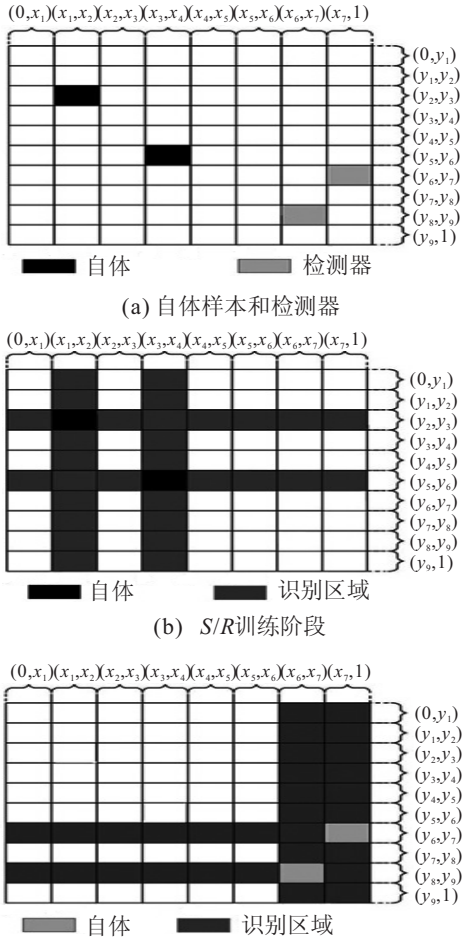
1) 对于离散型或字符型属性, 将其按出现的次序依次映射到相应的邻域中, 假设特征向量的一个字符型属性存在 m 种字符, 则将该属性对应的空间划分为 m 个邻域, 每种字符为一个邻域.

2) 对于连续型属性, 首先将该属性数据正规化到区间 $(0,1)_i$ 内, 即

$$x_i'=\frac{x_i-\min_i+\sigma}{\max_i-\min_i+\delta}. \tag{7}$$

其中: $\max_i$ 和 $\min_i$ 分别为该属性的最大与最小值, σ 和 δ 足够小并确保数据不包含顶点 0 和 1. 然后将数据映射到已经划分好的 m_i 个邻域中.

为了方便计算, 对形态空间的每一维按照划分的邻域数进行编码. 例如将第 i 维划分为 m_i 个邻域, 那么空间在第 i 维的编码是 $[0,m_i-1]$ 的一个自然数. 扩展到 n 维空间, 每个邻域的编码是一个 n 维向量 $X=(x_1,x_2,\cdots,x_n)$. 图 2(a) 表示二维邻域空间含 8×10 个互不相交的邻域, 每个自体/检测器是空间上的一个邻域, 且使用其顶点进行编码和标识. 其中: 黑色矩形代表自体, 其编码分别为 (1,2) 和 (3,5); 浅灰色矩形代表检测器, 其编码分别为 (7,6) 和 (6,8).



(a) 自体样本和检测器

(b) *S*/*R*训练阶段

(c) 检测阶段

图 2 二维邻域形态空间

3.2 邻域检测算法

在二进制空间和实值空间里, 检测算法以否定选择算法为基础衍生出了二进制否定选择算法 BNS (Binary NSA) 和实值否定选择算法 RNS (Real-valued NSA). 在邻域空间里, 检测算法可以相似地由否定选择算法衍生而得到邻域否定选择算法 NNS (Neighborhood NSA).

输入: N_d 为检测器要求数 (初始 n_d 为 0), $m[n]$ 为各维属性邻域划分数向量, ρ 为训练阈值;

输出: 检测器集 D.

```
begin
将各自体样本按m[n]正规化;
while (n_d < N_d) {
  随机生成候选邻域检测器 d; i = 0;
  if (H(d, s_i) > ρ, (i = 1, 2, ··· , N_S))
   {丢弃 d; Break;}
  else {D ← d; n_d + +}
}
end
```

匹配规则 $H(d,s_i)$ 定义为定义样本间相同维数

(即相似度), 并与预先设定的阈值 ρ 相比来训练检测器并检测异常. 在训练阶段, 对比候选检测器与每个自体样本, 如果存在相似度大于 ρ 的情况则消除候选检测器, 反之将候选检测器加入成熟检测器集合; 在检测阶段, 对比网络特征提取样本与每个检测器, 如果存在相似度大于 ρ 的情况则报警, 否则视该样本为正常. 有

$$H(X,Y)=\sum_{i=1}^{n}\overline{(x_i\oplus y_i)}. \tag{8}$$

$H(X,Y)$ 值越大表示 X 和 Y 越相似. 匹配规则可以简单描述为: 邻域 X 和 Y 匹配, 当且仅当他们存在 ρ 个相同的属性. 图 2(b) 显示了 NNS 在二维邻域形态空间下训练阶段的匹配过程, 其中训练阈值 ρ_{tr} 为 1. 空间内有两个自体, 通过匹配规则可得到与自体样本同行或同列的邻域都是自体样本的识别范围, 如果有候选检测器在这些范围内, 则将他们丢弃. 图 2(c) 显示了在二维邻域形态空间下检测阶段的匹配过程, 其中检测阈值 ρ_{de} 也为 1. 假设训练共生成 2 个检测器, 通过匹配规则可得到与检测器同行或同列的邻域都是检测器的识别范围, 在这些范围内存在的样本将被检测器检测为异常.

4　分析与实验

4.1　邻域表示法与实值表示法的比较

假设在 n 维邻域空间中, 第 i 维被分割成 m_i 个邻域, 则算法的搜索空间即形态空间中邻域的个数为 $N=\prod_{i=1}^{n}m_i$, 即 $N_d+N_R=N$, 其中 N_R 为非自体数, $N_d\leqslant N_R$. 经计算, NNS 与 BNS 的时间代价和空间代价相当. 邻域表示法较实值表示法的优势可以体现在以下几方面:

1) 邻域空间依然会出现 S 多分区问题, 但由于各邻域互不相交, 实值 S/R 边界交叉不会在邻域空间内出现. 另外, 每个邻域内的数据属于同一类型, 这就避免了实值空间单点检测器造成的样本高重叠率问题和噪声问题.

2) 邻域空间用不同的处理方式将各种类型的数据归一化成同一种类型的数据——邻域. 该方式克服了不同类型数据对于算法检测结果的影响, 同时也避免了 RNS 简单地将不同值域的属性正规化到同一尺度, 破坏了数据原有的尺度. 同时采用 Minkowski 体系距离的匹配规则不能正确衡量匹配度, 会在结果中造成一定的偏颇等问题.

3) NNS 匹配规则与 RNS 采用的 Minkowski 体系距离匹配规则相比, 由于摒弃了以识别体积为基础的匹配度策略, 避免了实值空间的维数灾难. 同时, NNS 的匹配规则削弱了样本属性间的耦合关系. 与 RNS 相比, 这种松散的耦合关系使得算法无需考虑数据属性间关系, 可以单独分析数据每维的特征, 简化了处理过程.

4.2　实验分析

实验选用 KDD Cup 1999 数据集一个 10 % 的子集进行测试[9], 该数据集包含 32 个连续型属性和 9 个离散型属性. 实验以数据集中的正常记录为自体集训练检测器, 并使用这些检测器检测数据集中的异常. 按照 NNS 正规化方法将数据归一化到形态空间中的邻域内. 并从以下几个方面分析邻域形态空间和 NNS 的性能:

1) 邻域数和自体样本数对于 NNS 性能的影响. 实验使用 100 % 和 50 % 的正常样本作为自体集训练检测器, 设定 $\rho_{tr}=\rho_{de}=6$, 各生成 5 000 个检测器, 结果如图 3 所示. 50 % 正常样本训练检测器时算法的检测率和误报率均比 100 % 训练时高, 且采用 100 % 训练时误报率一直为 0. 这说明随着自体数的减少, 检测器的识别范围增大, 检测率和误报率也随之增大, 所以自体集样本数是决定 NNS 综合性能的参数之一. 从图 3 还可以看出, 邻域数的增加使得算法的检测率和误报率均逐渐降低, 这说明邻域数的增加导致形态空间 (即算法的搜索范围) 的扩大. 在检测器数不变的情况下, 检测器非自体空间覆盖率降低, 造成算法性能的降低, 所以邻域数是另一个决定 NNS 综合性能的参数之一.

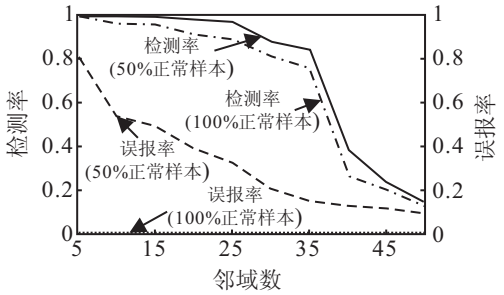
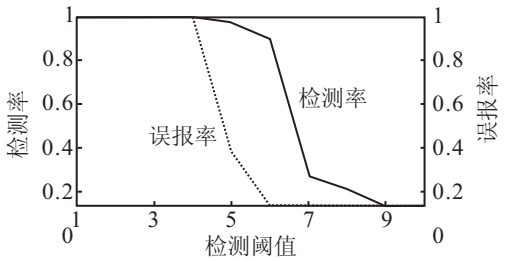


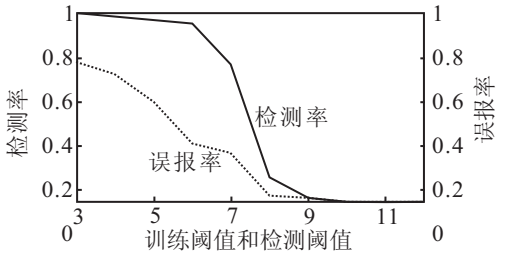
图 3　邻域数和自体样本数对 NNS 算法的影响

2) 训练阈值和检测阈值对于 NNS 性能的影响. 采用 100 % 的正常样本训练检测器, 设定属性邻域数为 25, 生成 5 000 个检测器. 首先固定 $\rho_{tr}=6$, 初始 $\rho_{de}=1$ 并逐渐增大, 结果如图 4(a) 所示. 当 $\rho_{de}<5$ 时算法的检测率均为 100 %, 但误报率较高. 当 $\rho_{tr}=\rho_{de}=6$ 时, 虽然检测率有所下降, 但误报率却降为 0. 继续增大 ρ_{de} 时, 虽然会使误报率保持在一个较低的水平, 但也会使检测率急剧下降. 所以当 $\rho_{tr}=\rho_{de}$ 时, 算法综合效果较好. 使用 50 % 的正常样本训练检测器, 初始令 $\rho_{tr}=\rho_{de}=3$, 以步长 1 同时增大其取值测试算法性能, 结果如图 4(b) 所示. 随着取值的增大, 检测率和误报率均有所下降, 当取值大于 10 时, 算法的检测

率和误报率几乎为0. 这说明同时增大两个阈值会减少检测器的识别范围, 降低检测器的识别能力, 从而造成算法性能下降.



(a) 训练阈值不等于检测阈值



(b) 训练阈值等于检测阈值

图 4 训练阈值和检测阈值对性能的影响

3) NNS 和 RNS 性能对比. 采用100 %的正常样本训练检测器, 前两种方法采用 RNS 和目前效果较好的改进算法 Boundary-aware RNS (BRNS)[10], 半径均为0.05; 后一种为 NNS, 设定属性邻域数为 25, $\rho_{tr} = \rho_{de} = 6$. 各生成 5 000 个检测器并检测异常 (进行 10 次取均值 Mean 与标准差 SD), 结果如表 1 所示. 从表 1 可以看出, RNS 效果最差, BRNS 虽比 RNS 效果好, 但其检测率的标准差远大于 NNS, 这表明邻域形态空间及 NNS 的检测性能优于实值空间和 RNS 及其改进算法.

表 1 3种算法检测结果 %

算法	检测率		误报率	
	MEAN	SD	MEAN	SD
RNS	27.99	0.125	0	0
BRNS	81.19	30.49	1.46	0.07
NNS	91.25	5.45	0	0

5 结 论

通过探讨实值表示法的弊端, 提出了邻域形态空间, 并给出了邻域检测算法 NNS. 邻域空间对免疫细胞表位组织形式进行建模, 采用离散拓扑理论设计自体/检测器, 并通过 NNS 进行检测. 实验结果表明, 与以往形态空间及相关算法相比, 邻域表示法及 NNS 具有定义简单、检测效果好、使用范围广等特点, 为免疫入侵检测技术及其他相关研究提供了一个新的研究方法. 下一步的研究方向是对于自体数、训练阈值和检测阈值等的设定.

参考文献(References)

[1] Dal D, Abraham S, Abraham A, et al. Evolutionary induced secondary immunity: An artificial immune systems based intrusion detection systems[C]. Proc of 7th Computer Information Systems and Industrial Management Applications. Ostrava: Czech Republic, 2008: 65-70.

[2] Dasgupta D, Gonzalez F. An immunity based technique to characterize intrusions in computer network[J]. IEEE Trans on Evolutionary Computation, 2002, 6(3): 281-291.

[3] Boukerche A, Machado R B, Juca R L. An agent based and biological inspired real-time itrusion detection and security model for computer network operations[J]. Computer Communications, 2007, 30(13): 2649-2660.

[4] Forrest S, Perelson S, Allen L. Self-nonself discrimination in a computer[C]. Proc of IEEE Society Symposium on Research in Security and Privacy and Privacy. Massachusetts, 1994: 202-212.

[5] Gonzalez F, Dasgupta D, Kozma R. Combining negative selection and classification technique for anomaly detection[C]. Proc of Congress on Evolutionary Computation. Honolulu, 2002: 705-710.

[6] Castro N, Timmis J. Artificial immune system: A new Computational Intelligence approach[M]. Berlin: Springer, 2002: 23-50.

[7] Kim J, Bentley P J. An evaluating negative selection in an artificial immune system for network intrusion detection[C]. Proc of Genetic and Evolutionary Computation Conf. San Francisco, 2001: 1330-1337.

[8] Greensmith J, Twycross J, Aickelin U. Dendritic cells for anomaly detection[C]. Proc of IEEE Congress on Evolutionary Computation. Vancouver, 2006: 664-671.

[9] UCIKDD Archive. KDD Cup 1999 Data[DB/OL]. (1999-10-28) [2006-10-01]. http: //kdd. ics. uci. edu/databases/kddcup99/ kddcup99.html.

[10] Zhou J, Dasgupta D. Applicability issues of the real-valued negative selection algorithm[C]. Proc of GEC'06. New York: ACM Press, 2006: 111-118.