一种基于抗原软子空间聚类的否定选择算法*

刘正军[1], 高江锦[2†], 杨 韬[1,2]
(1. 四川大学 计算机学院, 成都 610065; 2. 西华师范大学 教育信息技术中心, 四川 南充 637002)

摘 要: 否定选择算法(NSA)是免疫检测器生成的重要算法。传统否定选择算法在亲和力计算过程中未考虑不同种类抗原关键特征与冗余特征之间的差异性,存在算法检测性能较低的问题。对此,提出了一种基于抗原软子空间聚类的否定选择算法(ASSC-NSA)。该算法首先利用抗原软子空间聚类计算出不同种类抗原的各个关键特征及其权值,然后通过这些关键特征引导检测器生成以有效地减少冗余特征的影响,从而提高算法检测性能。实验结果表明,在 BCW 与 KDDCup 数据集上,相对于经典的否定选择算法,ASSC-NSA 能在误报率无明显变化的情况下显著地提高检测率。

关键词: 否定选择算法; 软子空间聚类; 异常检测
中图分类号: TP309.5; TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2018)03-0680-05
doi:10.3969/j.issn.1001-3695.2018.03.009

Improved negative selection algorithm based on antigen soft subspace clustering

Liu Zhengjun[1], Gao Jiangjin[2†], Yang Tao[1,2]
(1. *College of Computer Science, Sichuan University, Chengdu* 610065, *China*; 2. *Education Information Technology Center, China West Normal University, Nanchong Sichuan* 637002, *China*)

Abstract: Negative selection algorithm(NSA) is an important method of detector-generation. Traditional NSAs ignored the difference of key characteristic and redundant characteristic of different kinds of antigens in the process of affinity-computing, which led to the poor performance. To solve this problem, this paper proposed an improved negative selection algorithm based on antigen soft subspace clustering(ASSC-NSA). First, by utilizing the antigen soft subspace clustering algorithm, ASSC-NSA found out all key characteristics and their weights of different kinds of antigens. Then, using the key characteristics to guide the detectors generation, thus it could eliminate the adverse influence of redundant characteristics and improve the detection rate. Compared with classical NSAs, the experimental result on BCW and KDDCup data set shows that ASSC-NSA improves the detection rate significantly with the similar false alarm rate.

Key words: negative selection algorithm; soft subspace clustering; anomaly detection

0 引言

近年来,免疫计算(immune computation,IC)作为一种重要的智能计算方法得到了科学界的普遍关注[1~3]。由于免疫计算具有良好的鲁棒性、自适应、学习认知等特点,所以它被广泛运用在网络安全、组合优化、机器学习等[4~7]领域。否定选择算法(negative selection algorithm,NSA)作为 IC 的基础算法得到人们的高度重视,直接关系到 IC 的性能。Forrest 等人[8]提出的经典 NSA 算法模拟了生物免疫系统中免疫细胞成熟过程中的免疫耐受,通过清除能检测到自体的候选检测器减少对自体的识别,以达到对非自体的有效识别。为改进 NSA 的使用范围和性能,Gonzalez 等人[9]提出了实值否定选择算法(RNSA),该算法采用$[0,1]^n$实值空间编码抗原和抗体,并使用 Minkowski 距离计算抗原与抗体的亲和程度,将传统的 NSA 从二值逻辑推广到实值空间。Ji 等人[10]提出了可变半径的实值否定选择算法(RNSA with variable-sized detector,V-Detector),通过计算候选检测器中心与最近邻自体间的距离动态确定生成检测器的半径,有效地减少了检测器数量。Gong 等人[11]通过在训练过程中加入深入训练策略,生成自体检测器代替自体,以提高自体覆盖率和检测器生成效率。陈文等人[12]利用层次聚类方法对自体训练集进行聚类,以聚类中心取代自体元素进行自体耐受,进而减少检测器的生成时间代价。Zeng 和 Li 等人[13,14]通过动态设置不同的自体半径,使自体区域和非自体区域更容易被区分,从而提高检测率,降低误报率。

传统否定选择算法忽略不同种类样本各个特征之间的差异性,并未区分关键特征与冗余特征,无差别地通过所有特征属性来进行检测器与抗原的亲和度计算,大量冗余的特征(冗余特征包含相关性低的特征)降低了亲和度计算的精度,使不同类别的样本更难区分,从而导致算法检测率低。为解决以上问题,本文提出一种基于抗原软子空间聚类的否定选择算法(ASSC-NSA)。该算法首先利用抗原软子空间聚类算法计算出不同种类样本各个关键特征以及其重要权值,然后在各个关键特征组成的子空间利用特征权值引导检测器的生成,从而减

收稿日期: 2016-11-18; **修回日期**: 2017-01-04 **基金项目**: 国家自然科学基金资助项目(61572334);国家重点研发计划资助项目(2016YFB0800604);南充市应用技术研究与开发资金资助项目(16YFZJ0011)

作者简介:刘正军(1987-),男,四川成都人,硕士,主要研究方向为人工免疫、网络安全;高江锦(1977-),女(通信作者),四川南充人,副教授,硕士研究生,主要研究方向为网络安全、机器学习(gjj_cwnu@163.com);杨韬(1982-),男,四川遂宁人,副教授,硕士,主要研究方向为信息安全、人工免疫.

少冗余特征带来的不良影响,提高算法检测率。

1　基本定义

否定选择算法(NSA)通过删除识别自体(self)的随机检测器(detector),以达到有效地识别非自体(nonself)。其基本定义如下:

定义 1　抗原集 $Ag = \{ \cup_{i=1}^{n} ag_i \mid ag_i = (x_1, x_2, \cdots, x_i, \cdots, x_d), x_i \in [0,1] \}$,其中 d 为样本数据的维度;n 为样本空间的大小;x_i 为样本数据 ag_i 规范化后第 i 的数据;Ag 表示问题空间所有样本。

定义 2　自体集合 self$\subset Ag$,表示抗原集合中的正常样本,非自体集合 nonself = Ag-self,表示抗原集合中异常样本;$r_s \in R^+$ 为自体半径,被自体覆盖的区域为自体区域,其余为非自体区域。

定义 3　传统否定选择算法的检测器 $dt = \{ ab, r \mid ab \in Ag, r \in R^+ \}$,$ab$为检测器的中心位置向量;$r$ 为检测器半径,与检测器的距离小于 r 的抗原被判断为非自体抗原。

2　ASSC-NSA 算法实现策略

2.1　问题描述及理论分析

抗原之间的差异往往是由若干个关键的特征所引起的,不同类的抗原分布在不同关键特征组成的子空间(图 1)。随着维度的增加,抗原样本数据集存在以下两个特点[15]:a)样本分布较稀疏,不同种类的样本集中在不同的低维子空间内,图 1 所示的情况更普遍;b)高维空间中两点之间的距离趋于相同,一个样本点到其最近的邻居点和最远的邻居点的距离趋于相等,基于全局所有特征的距离更难区分样本之间的差异。因此否定选择算法不区分不同种类抗原各个特征的差异性(平等对待关键特征和冗余特征)的亲和力计算方法,会使冗余特征对亲和度计算造成不良影响。例如,算法随机生成一个点,由于冗余特征的影响,此点到自体和非自体的亲和度(距离)趋于相同的值,难以有效地区分自体和非自体。

图 1 展示了一个含 600 个抗原样本的数据集,数据集有三个属性特征,共有红($x = 0.4 \sim 0.6, y = 0.3 \sim 0.5, z = 0 \sim 1$)、蓝($x = 0.2 \sim 0.4, y = 0 \sim 1, z = 0.3 \sim 0.5$)、黑($x = 0 \sim 1, y = 0.5 \sim 0.7, z = 0.5 \sim 0.7$)三类数据(见电子版),每类数据集中分布在较关键的两个属性特征组成的子空间中。从图 1 中可以看出,红色数据的关键特征为 x、y;蓝色数据的关键特征为 x、z;黑色数据的关键特征为 y、z。

定理 1　n 维样本点 a、b、c,其中 a、b 属于同类别 X,c 属于类别 Y,通过对样本特征进行加权可使同种类样本的距离小于异类样本的距离,即 $\mathrm{dist}(a,b) < \mathrm{dist}(b,c)$、$\mathrm{dist}(a,b) < \mathrm{dist}(a,c)$。

证明　假设类 X 的关键特征有 m 个,其他特征为冗余特征,将样本的特征重新排列,前 m 个为关键特征,剩余的为冗余特征。极端情况下,设冗余特征权值为 0,则类 X 的样本特征权值向量为 $W_X = [k_i]_n = (k_1, k_2, \cdots, k_m, 0, \cdots, 0, \cdots, 0)$,其中 $\sum_{i=1}^{m} k_i = 1$,则 $\mathrm{dist}(a,b) = \sum_{i=1}^{m} k_i \times (a_i - b_i)^2 + \sum_{i=m+1}^{n} 0 \times (a_i - b_i)^2$,$\mathrm{dist}(b,c) = \sum_{i=1}^{m} k_i \times (b_i - c_i)^2 + \sum_{i=m+1}^{n} 0 \times (b_i - c_i)^2$。因为 a、b 属于同类 X,所以 a、b 在 m 个关键特征上的取值趋于相同,则 $\sum_{i=1}^{m} k_i (a_i - b_i)^2 \to 0$,而 b、c 不属于同一类,b、c 在 m 个关键特征上的取值必定不同,其值远大于 0,所以 $\sum_{i=1}^{m} k_i (a_i - b_i)^2 < \sum_{i=1}^{m} k_i \times (b_i - c_i)^2$,则 $\mathrm{dist}(a,b) < \mathrm{dist}(b,c)$,同理可证 $\mathrm{dist}(a,b) < \mathrm{dist}(a,c)$。

基于定理 1,提出基于权值的距离计算公式:

$$W\mathrm{dist}(dt,a) = \sqrt{\sum_{k=1}^{d} w_k^2 (dt_k - a_k)^2} \tag{1}$$

式(1)表示抗原样本 a 到检测器 dt 的距离,其中:a_k、dt_k 分别表示抗原样本和检测器的第 k 个特征的值;d 表示样本数据的维度;w_k 表示检测器 dt 的第 k 个特征值的权值。

定义 4　具有权值的检测器 $dt = \{ ab, w, r \mid ab \in Ag, w = (w_1, w_2, \cdots, w_d), r \in R^+ \cdots w_i \in [0,1], \sum_{i=1}^{d} w_i = 1 \}$,其中:$w_i$ 表示检测器第 i 维的权值;d 表示样本数据的维度;ab 表示检测器的中心向量;r 表示检测器的半径,与检测器的距离小于 r 的抗原被判断为非自体抗原。

样本之间的距离如表 1 所示。

表 1　样本之间的距离

样本点	类别	全部属性欧拉距离(dist)			子空间	子空间权值	子空间上的距离(Wdist)		
		a	b	c			a	b	c
a	红	0	0.6	0.54	x,y	0.5,0.5,0	0	0	0.1
b	红	0.6	0	0.22	x,y	0.5,0.5,0	0	0	0.1
c	蓝	0.54	0.22	0	x,z	0.5,0,0.5	0.27	0.11	0

抗原之间的差异由关键特征决定。平等对待抗原所有特征,使冗余特征模糊了不同类别之间的区别。例如图 1 中的 a、b、c 三点:$a = (0.5, 0.4, 0.9)$、$b = (0.5, 0.4, 0.3)$ 两个点属于红色类别,$c = (0.3, 0.4, 0.4)$ 点属于蓝色类别。从表 1 可知,未区分特征的重要性,通过抗原全部特征属性计算的欧拉距离:$\mathrm{dist}(a,b) = 0.6$、$\mathrm{dist}(a,c) = 0.54$、$\mathrm{dist}(a,c) = 0.22$,同类别 a、b 两点的距离大于不同类两点 a、c 和 b、c 的距离,抗原样本难以区分。

如果找出样本数据存在的各个子空间,并对其各个特征的重要性进行加权,挑选出最重要的特征子集,就能有效地减少或消除冗余特征的影响,从而提高分类的精确度。定理 1 表明,对样本特征进行加权可有效使不同样本之间的差异变得更明显,从而更容易被区分。本文通过抗原软子空间聚类得到由关键特征所形成的子空间以及其对应的特征权值。基于定理 1 提出距离计算式(1),并在否定选择算法中提出具有权值的检测器表示方式(定义 4)。由表 1 可知,可以得出红色点存在的子空间的权值为(0.5,0.5,0)、蓝色点存在的子空间的权值(0.5,0,0.5),根据式(1)重新计算 a、b、c 三点距离可得,同类两点 a、b 的距离为 0,异类两点 c 到 a 和 b 的距离均为 0.1,b 到 c 的距离为 0.11,a 到 c 的距离为 0.27。新的计算结果表明,不同类别点的距离大于同类别点的距离,样本能被正确区分。

2.2　ASSC-NSA 算法

ASSC-NSA 算法如算法 1 所示。其主要分为两个主要部分:a)(第 1 ~ 3 步)初始检测器生成阶段,通过抗原软子空间聚类算法(subspace-clustering)对非自体和自体进行聚类选出非自体不同类别抗原的聚类及其各特征的权值,根据定义 4 组建对应的初始检测器并加入检测器集合,每个初始检测器代表一类子空间;b)(第 4 步)带特征权值的随机检测器生成阶段,基于初始检测器,通过检测器生成函数(detector-generating)生成更多经过自体耐受的并带有特征权值的随机检测器。

算法 1 ASSC-NSA 算法

ASSC-NSA(self,nonself,p,C,detector):

输入:自体训练集 self,非自体训练集 nonself,期望覆盖率 p,初始聚类数量 c。

输出:检测器集合 detector。

1 令 detector = Φ,antigens = [nonself;self];

2 调用 subspace-clustering(antigens,c)得到非自体聚类的位置矩阵 Z_N、权值矩阵 W_N、半径矩阵 R_N;

3 将$\langle z_i, w_i, r_i\rangle$组成的初始检测器加入 detector,其中 z_i、w_i、r_i 分别是 Z_N、W_N、R_N 的第 i 个元素;

4 基于初始检测器,利用检测器生成算法 detector-generating(detector,self,p)生成随机检测器集合,并加入到检测器集合 detector。

2.2.1 抗原软子空间聚类算法

抗原软子空间聚类算法的主要目的是找不同抗原的子空间及其各特征的权值,为生成初始检测器作准备。其核心思想是找出使类内距(J_{SC})最小[16]的归属矩阵、位置矩阵以及权值矩阵。表 2 列出了算法中常用的符号。通过拉格朗日乘子法,可以得到当类内距(J_{SC})最小时,归属矩阵 U、位置矩阵 Z、权值矩阵 W 的更新公式。其中聚类的权值矩阵代表特征对于聚类的重要程度,其大小可用来判断此特征是否冗余。抗原软子空间聚类的具体算法如算法 2 所示。

表 2 算法常用的符号

符号	含义	取值
c	聚类的数量	由具体数据集决定
d	样本数据特征总个数	由具体数据集决定
n	样本数据大小	由具体数据集决定
x_{jk}	第 j 样本的第 k 特征的值 $j\in n$ and $k\in d$	由具体数据集决定
J_{SC}	目标函数(类内距)	式(2)
$U=[u_{ij}]_{c\times d}$	样本的归属矩阵	式(3)
$Z=[z_{ij}]_{c\times d}$	聚类中心的位置矩阵	式(4)
$R=[r_i]_c$	聚类的半径矩阵	式(5)
$W=[w_{ij}]_{c\times d}$	聚类的权值矩阵	式(6)
$B=[b_i]_c$	聚类的类别矩阵	式(7)

算法 2 抗原软子空间聚类算法

subspace-clustering(antigens,c):

输入:训练集合 antigens,聚类个数 c。

输出:聚类中心位置矩阵 Z、权值矩阵 W、半径矩阵 R。

1 在训练集合 antigens 中随机选取 c 个样本点作为初始聚类中心,令 $W=[1/D]_{C\times D}$,迭代次数 $s=1$,Z 为选中的 c 个样本点的特征值,聚类的类别 $B=[-1]_c$;

repeat:

2 根据式(3)更新 antigens 的归属矩阵 U;

3 根据式(4)更新聚类中心位置矩阵 Z;

4 根据式(6)更新权值矩阵 W,$s=s+1$;

until ($\|Z^s-Z^{s-1}\|_\infty<\delta$ 且 $\|W^s-W^{s-1}\|_\infty<\delta$)

5 根据式(7)计算 B,选出类别为 1 的非自体聚类对应的 Z、W、R;

6 运用式(5)计算聚类半径矩阵 R;

7 计算 antigens 中自体到聚类 i 最小距离 r_m,如果 $r_m\leqslant r_i$,则 $r_i=r_m$。

其中:δ 是一个很小的正数,用来控制算法中循环的终止条件。例如 $\delta=10^{-6}$,当前后两次的聚类中心位置矩阵的无限范数和前后两次的权值矩阵的无限范数都小于 δ 时循环终止。Z^s、W^s 分别表示第 s 次迭代时 Z、W 的值。c 为初始聚类个数。

聚类算法目标函数:J_{SC},其中:ε 是避免属性零散度时出现除零错误而引入的非常小的一个常数;β 是模糊因子、w_{ik}^{β} 是对各个属性上样本与聚类中心的距离进行模糊加权。模糊加权 w_{ik}^{β} 可以看做是经典加权 w_{ik} 的推广。一般令模糊指数 $\beta>1$。当 $\beta\to1$ 时,w_{ik}^{β} 接近于经典加权 w_{ik}。

$$J_{SC}=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}\sum_{k=1}^{d}w_{ik}^{\beta}(x_{jk}-z_{ik})^2+\varepsilon\sum_{i=1}^{c}\sum_{k=1}^{d}w_{ik}^{\beta}$$
$$\text{s.t.}\quad u_{ij}\in\{0,1\},0\leqslant w_{ik}\leqslant1 \text{ and } \sum_{k=1}^{d}w_{ik}=1 \tag{2}$$

归属矩阵:$U=[u_{ij}]_{c\times d}$,每个数据样本必须归属且只能归属一个聚类,当且仅当样本数据 x_j 与聚类中心的位置 z_i 的距离方差最小时,样本数据 x_j 隶属于聚类 i。

$$u_{ij}=\begin{cases}1 & i=\arg\min\limits_{m=1,\cdots,c}\sum\limits_{k=1}^{d}w_{mk}^{\beta}(x_{jk}-z_{mk})^2\\ 0 & \text{else}\end{cases} \tag{3}$$

位置矩阵:$Z=[z_{ik}]_{c\times d}$,聚类中心的位置 z_{ik} 为所有隶属于第 i 聚类的样本数据的第 k 维的均值。

$$z_{ik}=\frac{\sum\limits_{j=1}^{n}u_{ij}x_{jk}}{\sum\limits_{j=1}^{n}u_{ij}} \tag{4}$$

半径矩阵:$R=[r_i]_c$,第 i 个聚类的半径 r_i 为所有归属于此聚类的样本数据与此聚类中心的距离的平均值。

$$r_i=\frac{\sum\limits_{j=1}^{n}u_{ij}\sqrt{\sum\limits_{k=1}^{d}w_{ik}^2(x_{jk}-z_{ik})^2}}{\sum\limits_{j=1}^{n}u_{ij}} \tag{5}$$

权值矩阵:$W=[w_{ik}]_{c\times d}$,权值 w_{ik} 表示第 k 维特征对第 i 个聚类的聚类贡献度,ε 是避免属性零散度时出现除零错误而引入的非常小的一个常数。

$$w_{ik}=\left(\sum_{l=1}^{d}\left[\frac{\sum\limits_{j=1}^{n}u_{ij}(x_{jk}-z_{ik})^2+\varepsilon}{\sum\limits_{j=1}^{n}u_{ij}(x_{jl}-z_{il})^2+\varepsilon}\right]^{1/(\beta-1)}\right)^{-1} \tag{6}$$

类别矩阵:$B=[b_i]_c$,$b_i=1$ 表示此聚类为非自体聚类,否则为自体聚类;∂ 为类别控制阈值,类别 i 含非自体比例超过 ∂ 表示类 i 为非自体聚类。

$$b_i=\begin{cases}1 & \dfrac{\sum\limits_{j=1}^{n}u_{ij}f(x_j)}{\sum\limits_{j=1}^{n}u_{ij}}>\partial\\ 0 & \text{else}\end{cases},\quad f(x)=\begin{cases}1 & x\in\text{nonself}\\ 0 & x\in\text{self}\end{cases} \tag{7}$$

2.2.2 检测器生成算法

检测器生成算法如算法 3 所示。基于抗原软子空间聚类得到的初始检测器,筛选出符合条件的候选检测器并计算其权值与半径。每个初始检测器各自代表一种类型的抗原,其权值代表这类抗原所在的子空间。如果随机生成的点 a 到初始检测器 dt_i 的距离 $W\text{dist}(dt_i,a)$ 最小,表明点 a 最有可能在检测器 dt_i 代表的子空间内。所以,本文设置候选检测器的权值为离其最近的一个初始检测器的权值。根据抗体耐受的原理,候选检测器的半径为自体到候选检测器的最小距离(基于权值的距离)。图 2 展示候选检测器 dt 的权值和半径的计算方法。其中,红色点代表自体,蓝色实线圆表示初始检测器,黑色实线圆表示候选检测器 dt(见电子版)。a 为候选检测器 dt 的位置向量;离候选检测器 dt 最近的初始检测器是 dt_1,所以 dt 的权值是 $dt_1\times w$;离 dt 最近的自体为 e,所以 dt 的半径 $r=|ae|=W\text{dist}(d,e)$。

算法 3 检测器生成算法

detector-generating(detector,self,p)

输入:初始检测器集合 detector,自体集合 self,检测器期望覆盖率 p。

输出:检测器集合 detector。

repeat:

1　在取值范围内随机生成一个候选检测器的位置向量 x。

2　应用式(1)计算出 x 到初始检测器的距离,将距离 x 最近的初始检测器的权值 w 作为 x 的权值。如果 x 在检测器的范围内,递增检测器重复数量计数器 m。

3　应用式(1)计算 x 与自体的最短距离 r。如果 x 不在自体的半径范围内,则递增检测器个数计数器 t;否则删除 x 跳转第 1 步。

4　将$\langle x,w,r\rangle$组成的检测器加入 detector,运用式(8)计算终止条件 $T=\text{coverage}(p,t,m)$。如果 $T=-1$,则令 $t=m=0$。

until $T=1$

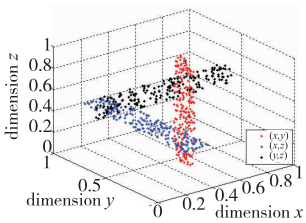

图1　抗原样本数据分布

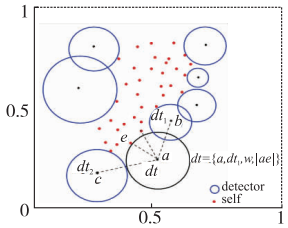

图2　候选检测器权值及半径的计算方法

本文使用统计学中的假设检验方法[10]计算终止条件,如式(8)所示,当结果为 -1 时,表示需清空计数器并继续循环;当结果为 0 时,继续循环产生检测器;当结果为 1 时,表示算法达到期望覆盖率,终止算法。Z_a 是很小的控制常数,如 $Z_a=0.001$,$G>\max(5/p,5/(1-p))$。

$$\text{coverage}(p,t,m)=\begin{cases}-1 & \text{if } t\geqslant G\\ 0 & \text{else if } \dfrac{m}{\sqrt{t\times p\times(1-p)}}-\sqrt{\dfrac{t\times p}{1-p}}<Z_a\\ 1 & \text{else}\end{cases}\tag{8}$$

3　实验结果与分析

3.1　实验设置

本文通过在 UCI(University of California Irvine)的 breast cancer wisconsin(BCW)和 KDDCup99 两组经典实验数据集[17]进行实验,以验证 ASSC-NSA 算法的性能,其中 BCW 是威斯康星州病人的胸腺癌特征数据集,KDDCup99 是美国国防部高级规划署(DARPA)在 MIT 林肯实验室进行的一项入侵检测所提取的网络流量数据,这些数据集被广泛地应用在机器学习、异常检测、人工免疫检测器的生成[2,11~14]。详细的数据集描述见以下实验。

本文采用的对比算法是 RNSA(代表经典的固定半径 NSA)、V-Detector(VD)(代表经典的可变半径 NSA)。实验采用检测率 DR(detection rate)和误报率 FAR(false alarm rate)对本算法产生的检测器性能进行衡量。DR 和 FAR 的定义公式如下:

$$DR=TP/(FN+TP),FAR=FP/(FP+TN)$$

其中:TP、TN、FP、FN 分别表示正确肯定的非自体、正确否定的自体、错误肯定的自体、错误否定的非自体。

3.2　算法参数设置

不同的参数以及策略将影响算法的性能,本文在 BCW 数据集上测试不同参数对算法性能的影响。BCW 数据集有 458 条正常数据和 241 条异常数据,其中“2”代表正常的数据(self),“4”代表异常的数据(nonself);每条数据共有 10 个维度,最后一维为类别特征。本文将除类别特征外的其他特征归一到形态空间$[0,1]^9$。

自体半径和检测器期望覆盖率是影响算法性能的两个重要参数。在 BCW 数据集中,本实验随机选取 50% 的自体样本以及非自体样本作为训练集,将其余的样本作为测试集。测试的自体半径为[0.01,0.6],检测器期望覆盖率为[90%,99.99%]。经实验得到聚类的最佳个数为 60,算法终止条件参考式(8)。自体半径与检测器期望覆盖率对算法性能影响的实验结果如图 3 和 4 所示。从图中可以看出,算法的检测率 DR 和误报率 FAR 随着自体半径的增加而减少,同时随着检测器期望覆盖率的增加而增加,当自体半径为 0.2、期望覆盖率为 99% 时检测效果相对较好。因为自体半径的增加导致可被检测到的自体区域增加、非自体区域减少,而覆盖率的减少导致可被检测到的非自体区域减少、自体区域增加,从而使检测率和误报率变得更低。

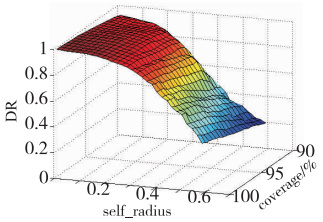

图3　期望覆盖率和自体半径对检测率的影响

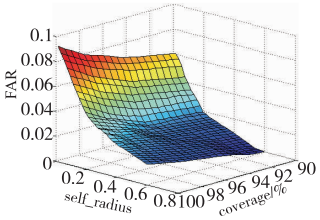

图4　期望覆盖率和自体半径对误报率的影响

初始聚类的数量是另外一个影响算法性能的重要参数。此次实验随机选取 50% 的自体样本以及非自体样本作为训练集合,将其余的样本作为测试集合。通过前一个实验可知,在 BCW 数据集上,当自体半径设置为 0.2、检测器期望覆盖率设置为 99% 时检测率较高、误报率很低,此时为最佳设置。重复 20 次后取平均值,结果如图 5 所示。当聚类的数量超过 60 以后,算法的检测率无明显升高,误报率无明显降低,但是算法的运行时间却随聚类的数量增加而急剧增加。因为随聚类的数量增加,将会增加大量的距离计算,预处理的时间将会显著增加,但算法性能趋于最佳,因而增加不明显,所以 BCW 数据集上聚类的最佳个数应为 60。

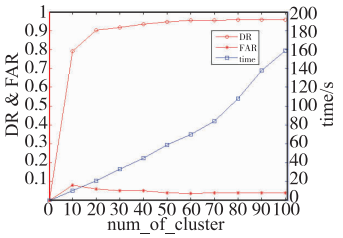

图5　初始聚类个数对算法性能的影响

因此,在 BCW 数据集上实验参数最佳设置是:自体半径为 0.2,检测器期望覆盖率为 99%,初始聚类个数为 60;同理,在 KDDCUP 上,实验参数最佳设置是:自体半径为 0.000 2,检测器期望覆盖率为 99%,初始聚类个数为 200。

3.3　对比实验

为了测试提出的算法的检测性能,本文在数据集 BCW 和 KDDCUP 数据集上进行对比实验。

BCW 数据集的介绍与实验参数设置参考 3.2 节,其中根据实验得到 RNSA 算法的检测器半径最佳为 0.1。图 6 和 7 显示的是相同参数设置下经过 20 次 BCW 数据集上重复实验的结果,ASSC-NSA 比 VD 和 RNSA 具有更高的检测率和更低的误报率。其统计数据如表 3 所示。

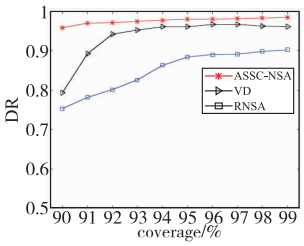

图6 三种算法在BCW上的检测率对比

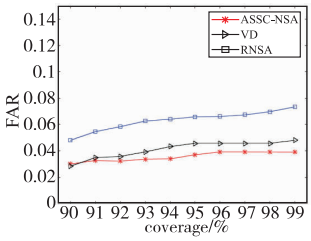

图7 三种算法在BCW上的误报率对比

表 3 算法在 BCW 和 KDDCup 数据集上的对比结果

算法	BCW		KDDCup	
	DR	FAR	DR	FAR
ASSC-NSA	0.979 7	0.036 7	0.938 8	0.042 6
VD	0.936 7	0.041 9	0.825 6	0.044 6
RNSA	0.848 1	0.062 0	0.746 9	0.051 9

KDDCUP 数据集有 Normal、DOS、R2L、U2R、Probing 五个类别，每个数据具有 42 个特征，最后一个是类别特征，其中有三个特征是字符串形式，按类别将其转换成不同的整数，最后整体归一化到[0,1][41]。KDDCUP 中非自体数据大部分是 DOS 攻击，并且很多 DOS 攻击是重复的数据。为避免选出的都是 DOS 攻击，本实验分别从 Normal、DOS、R2L、U2R、Probing 中按 10%、5%、50%、50%、50% 的比例随机选取数据作为训练集合，按同样比例选取测试数据集，将 Normal 数据作为自体，其他的作为非自体。为了测试相同条件下算法的性能对比，NSA 算法的终止条件由检测器总个数修改成检测器的期望覆盖率。图 8 和 9 显示在 KDDCUP 数据集上的对比实验结果。经实验可得自体半径的最佳设置为 0.000 2、NSA 的检测器的最佳半径为 3.6、最佳聚类个数 $C=200$。从图中可以看出，ASSC-NSA 比 RNSA 和 VD 具有更高的检测率和更低的误报率。其统计数据如表 3 所示。

以上实验表明，ASSC-NSA 具有更高的检测率和更低的误报率。表 3 是以上两个实验结果的平均值。由表 3 可知，当数据集特征的个数增加时，ASSC-NSA 算法的性能提升得更明显：在 41 维的 KDDCUP 数据集上，算法的平均检测率比 VD 和 RNSA 分别提高了 11.32% 和 19.17%；而在 9 维的 BCW 数据集上，检测率只分别提高了 4.30% 和 13.16%。因为当数据集特征的个数（维度）增加时，存在大量的冗余特征混淆了不同样本之间的区别；而且随着维度的增加，计算出的样本点之间的距离趋于相同，自体和非自体变得更加难以区分，ASSC-NSA 通过抗原软子空间聚类找到不同抗原的各个子空间以及其权值，区分出不同种类抗原的关键属性和冗余属性，让自体与非自体更易被分辨，所以能够提高检测率，降低误报率。

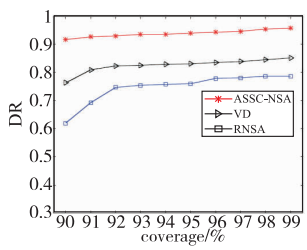

图8 三种算法在KDDCup上的检测率对比

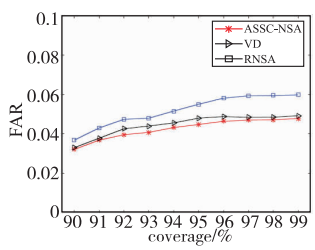

图9 三种算法在KDDCup上的误报率对比

4 结束语

针对否定选择算法忽略不同种类样本特征导致检测性能过低的问题，本文提出一种基于软子空间聚类的否定选择算法（ASSC-NSA）。算法通过抗原软子空间聚类得到不同种类样本的子空间及其特征的权值，在关键特征组成的子空间内生成检测器来引导随机检测器的生成，从而提高算法检测率，降低误报率。对比实验结果表明，在 BCW 和 KDDCup 两个数据集上 ASSC-NSA 算法比经典的否定选择算法具有更好的检测效果，特别是在高维度的数据集上检测率提升得更高。在后续工作中，将研究如何自动获取最佳的初始聚类的数量以及最佳的自体半径。

参考文献：

[1] Aragón V S, Esquivel S C, Coello C A C. An immune algorithm with power redistribution for solving economic dispatch problems[J]. Information Sciences, 2015,295(C):609-632.

[2] Li Dong, Liu Shulin, Zhang Hongli. Negative selection algorithm with constant detectors for anomaly detection[J]. Applied Soft Computing, 2015,36: 618-632.

[3] Gong Tao. High-precision immune computation for secure face recognition[J]. International Journal of Security and Its Applications, 2012, 6(2):293-297.

[4] Hart E, Timmis J. Application areas of AIS: the past, the present and the future[J]. Applied Soft Computing, 2008,8(1): 191-201.

[5] Silva G C, Palhares R M, Caminhas W M. Immune inspired fault detection and diagnosis: a fuzzy-based approach of the negative selection algorithm and participatory clustering[J]. Expert Systems with Applications, 2012,39(16):12474-12486.

[6] 李涛. 基于免疫的计算机病毒动态检测模型[J]. 中国科学,F辑:信息科学, 2009,39(4):422-430.

[7] Fakhari S N S, Ziabari M T. A self adaptive algorithm for classification based on negative selection technique[J]. Artificial Intelligence and Application, 2014,1(2).

[8] Forrest S, Perelson A S, Allen L, *et al*. Self-nonself discrimination in a computer[M]. [S.l.]:IEEE Computer Society,1994:202-212.

[9] Gonzalez F A, Dasgupta D, Nino L F. A randomized real-valued negative selection algorithm[C]//Proc of the 2nd International Conference on Artificial Immune Systems. 2003:261-272.

[10] Ji Zhou, Dasgupta D. V-detector: an efficient negative selection algorithm with probably adequate detector coverage[J]. Information Sciences, 2009,19(9):1390-1406.

[11] Gong Maoguo, Zhang Jian, Ma Jingjing, *et al*. An effcient negative selection algorithm with further training for anomaly detection[J]. Knowledge-Based Systems, 2012,30: 185-191.

[12] 陈文, 李涛, 刘晓洁, 等. 一种基于自体集层次聚类的否定选择算法[J]. 中国科学,F辑:信息科学, 2013,43(5): 611-625.

[13] Zeng Jinquan, Liu Xiaojie, Li Tao, *et al*. A self-adaptive negative selection algorithm used for anomaly detection[J]. Progress in Natural Science, 2009,19(2):261-266.

[14] Li Dong, Liu Shulin, Zhang Hongli. A negative selection algorithm with online adaptive learning under small samples for anomaly detection[J]. Neurocomputing, 2015,149(C): 515-525.

[15] Beyer K, Goldstein J, Ramakrishnan R, *et al*. When is "nearest neighbor" meaningful[C]//Proc of International Conference on Database Theory. Berlin: Springer, 1999: 217-235.

[16] Zhu Lin, Cao Longbing, Yang Jie, *et al*. Evolving soft subspace clustering[J]. Applied Soft Computing, 2014,14(1):210-228.

[17] UCI 数据集[EB/OL]. http://archive.ics.uci.edu/ml/datasets.html.