



## 基于数字微分的函数化树突状细胞算法模型

张 艺, 周 雯, 梁意文, 谭成予

(武汉大学 计算机学院, 武汉 430072)

**摘 要:** 树突状细胞算法(DCA)是一种模拟人体免疫系统中抗原提呈的算法,可以快速有效地将输入数据分为正常和异常,然而现有 DCA 模型普遍存在形式化描述不清晰且信号提取受人工经验影响的不足。为此,在 hDCA 模型的基础上,提出一种基于数字微分的函数化 DCA 模型。在预处理阶段引入数字微分方法,根据数据变化趋势自适应提取信号并随机动态采样抗原,去除对时序敏感的数据序列。在此基础上,对输入信号加以融合得到决策信号,并进行抗原背景环境分类处理。将 ndhDCA、DCA 和 hDCA 应用于 WBC 和 KDD99 数据集进行对比,实验结果表明,ndhDCA 对有序数据集和无序数据集均具有高准确率和低误报率,同时可降低输入数据顺序的敏感性。

**关键词:** 树突状细胞算法;hDCA 模型;数字微分;人工免疫系统;特征提取

开放科学(资源服务)标志码(OSID):



**中文引用格式:** 张艺,周雯,梁意文,等. 基于数字微分的函数化树突状细胞算法模型[J]. 计算机工程,2020,46(9): 54-60.

**英文引用格式:** ZHANG Yi, ZHOU Wen, LIANG Yiwen, et al. Model of functional dendritic cell algorithm based on numerical differentiation[J]. Computer Engineering, 2020, 46(9): 54-60.

### Model of Functional Dendritic Cell Algorithm Based on Numerical Differentiation

ZHANG Yi, ZHOU Wen, LIANG Yiwen, TAN Chengyu

(School of Computer, Wuhan University, Wuhan 430072, China)

**[Abstract]** The Dendritic Cell Algorithm (DCA) is an algorithm for simulating antigen presentation in the human immune system, which can divide input data into normal and abnormal data quickly and effectively. However, the existing DCA models are generally lack of clear formal description and their signal extraction is affected by artificial experience. To address the problems, this paper proposes a numerical differentiation-based functional dendritic cell model, named ndhDCA, by improving the hDCA model. In the preprocessing stage, the numerical differentiation method is introduced to extract the signal adaptively according to the trend of data change and to randomly and dynamically sample the antigen to remove the time-sensitive data sequence. On this basis, the input signal is fused to obtain the decision signal, and the antigen background environment is classified. ndhDCA, DCA and hDCA are compared on WBC and KDD99 data sets. The experimental results show that ndhDCA has higher accuracy and lower false positive rate in both ordered data sets and unordered data sets, and overcomes the sensitivity of the data sequence.

**[Key words]** Dendritic Cell Algorithm (DCA); hDCA model; numerical differentiation; Artificial Immune System (AIS); feature extraction

DOI:10.19678/j.issn.1000-3428.0055380

## 0 概述

人工免疫系统(Artificial Immune System, AIS)是受到人体免疫系统启发,模拟其行为、机制和功能的计算机系统<sup>[1]</sup>,可用于解决信息安全领域的多个问题,包括恶意代码检测、入侵检测、垃圾邮件检测、shellcode 检测等<sup>[2]</sup>。AIS 算法根据应用类别主要分

为优化算法、学习算法和异常检测算法。异常检测算法能够解决计算机安全的相关问题<sup>[3]</sup>,其中代表算法为树突状细胞算法(Dendritic Cell Algorithm, DCA)。

目前树突状细胞算法已经成功应用于诸多安全相关领域,解决了具体的入侵检测问题,其中包括端口扫描检测<sup>[4]</sup>、僵尸网络检测<sup>[5]</sup>、机器人安全分类器<sup>[6]</sup>、ping 扫描检测<sup>[7]</sup>、服务器攻击检测<sup>[8]</sup>以及电网

**基金项目:** 国家自然科学基金“计算机免疫智能的连续应答机制及其应用研究”(61877045)。

**作者简介:** 张 艺(1995—),女,硕士研究生,主研方向为计算机免疫;周 雯,博士;梁意文,教授、博士;谭成予,副教授、博士。

**收稿日期:** 2019-07-03 **修回日期:** 2019-08-19 **E-mail:** chitty\_zy@163.com

攻击检测<sup>[9]</sup>等,近年来也被应用于地震预测<sup>[10]</sup>和网络水军检测<sup>[11]</sup>等方面。

GREENSMITH 等人于 2005 年提出最初的树突状细胞算法 (prototype DCA, pDCA) 模型,并证明这种基于种群的算法能够对有序数据进行两类判别<sup>[12]</sup>,此后其又完善了算法使用的步骤和流程,提出了更完整的模型 (libtissue DCA, ltDCA)<sup>[13]</sup>。尽管 ltDCA 已被证明在诸多应用中有效,但其存在算法可控性差 (相互作用参数和随机元素多)、信号预处理方法不明确、算法缺乏形式化规范、可分析性差等不足<sup>[14]</sup>。针对可控性差的问题, GREENSMITH 等人通过去除大部分随机性参数,提出了轻量级的算法模型——确定性树突状细胞算法 (deterministic DCA, dDCA) 模型<sup>[15]</sup>。近年来也有学者对原算法的信号提取进行优化,如结合 XGBoost 的改进算法<sup>[16]</sup>和集成 PCA 的改进算法<sup>[17]</sup>,但都没有给出明确的形式化描述。在理论分析方面, GU 等人对 DCA 和 dDCA 进行了形式化描述<sup>[18]</sup>,但描述与实现细节混杂在一起,缺少数据流和函数定义; GREENSMITH 等人引入事件流的概念,在 dDCA 的基础上,对算法进行了函数化描述,形成了较为明确的规范 (deterministic DCA in Haskell, hDCA)<sup>[19]</sup>。在信号预处理阶段,通过特征选择机制提取信号的方式会损坏数据特征的隐含意义,因此, CHELLY 等人结合模糊粗糙集的方法<sup>[20]</sup>避免了信息损失,但该方法仍存在信号提取受人影响、普适性差的不足。

通过数字微分可以依据数据的变化和变化趋势来感知危险<sup>[21]</sup>,在信号提取阶段使用能够自适应提取危险信号 (DS) 和安全信号 (SS),解决上述算法普适性差的问题。为此,本文在 hDCA 模型的基础上,结合数字微分改进原算法中的信号提取过程,建立基于数字微分的函数化树突状细胞算法 (numerical differentiation based hDCA, ndhDCA) 模型,并将其应用于 UCI 机器学习库,与传统 DCA 和 dDCA 模型进行性能比较。

## 1 相关知识

### 1.1 数字微分

数据的变化及数据的变化趋势被认为是异常检测的基础。“危险”通常表现为数据的异常,可用于评估数据集中各项要素的重要性。数字微分<sup>[21]</sup>是一种通过研究离散数据集,分析数值变化和变化趋势来感知“危险”的方法,能够从全部特征集中提取应用环境的所需特征。数字微分包括数值微分、数组微分和向量微分,在实际使用中可根据数据的特点选用合适的算子并加以调整。

文献[22]利用数字微分从应用环境的信号中自适应地提取危险信号,建立基于变化的自适应危险信号提取模型。文献[23]将数字微分用于入侵检测

的数据预处理过程,提高了入侵检测的准确率。文献[24]将数字微分结合树突状细胞算法用于故障检测,在提高检测率的同时实现较早检测到渐变故障的目标。

本文借鉴数字微分的思想,将数据集的特征集定义为  $S = \{S_0, S_1, \dots, S_{n+1}\}$ , 将伴随其变化的一系列触发集合定义为  $T = \{T_0, T_1, \dots, T_{n+1}\}$ , 它们可以被看作是自变量和因变量。根据数据的变化,触发数据集会影响特征数据集。数字微分是一种计算因变量变化和变化趋势的方法,其特征变化量可以被定义为  $\Delta S_i = S_{i+1} - S_i (i = 0, 1, \dots, n)$ , 变化强度被定义为  $dS_i = \frac{\Delta S_i}{\Delta T_i}$ 。

### 1.2 DCA 发展及 hDCA 介绍

DCA 是一种免疫启发算法,最初基于天然树突细胞的功能,因为其具有非常低的 CPU 处理要求并且不需要大量的训练期,所以在应用于实时计算机安全问题时具有明显的优势。DCA 应用于端口扫描检测、僵尸网络检测、无线传感器不良行为检测、机器人安全等方面都取得了较好的效果。但由于算法中相互作用的参数数量过多,相关研究中参数设置和可变阈值设置较为随意。GREENSMITH 等人对 DCA 进行分解并测试其内在关系,对算法进行了参数简化,并提出新的异常度量,建立了可控性更强的确定性树突状细胞算法模型 dDCA<sup>[15]</sup>。2010 年, CHELLY 等人新增了两个信号模型,采用模糊集切换输入流数据,利用模糊子集修改信号变换方程,由此建立模糊树突状细胞模型 FDCM,提高了算法性能<sup>[25]</sup>。2011 年, GU 等人提出 xDCA<sup>[26]</sup>,通过引入自动化信号预处理过程,使用主成分分析法并利用特征选择机制自适应选择信号流的数据源。

在理论研究方面, OATES 等人于 2007 年发现了 DCA 算法的滤波和降噪属性,并指出算法具有线性分类器属性<sup>[27]</sup>。2009 年, STIBOR 等人使用点积构建信号处理阶段的模型,并且演示了其线性分类器属性,结果表明 DCA 可能遇到与线性分类器相同的问题:分类边界复杂造成错误和处理动态超平面问题的性能受损<sup>[28]</sup>。2011 年, GU 等人以支持向量机代替线性分类阶段,发现 DCA 的滤波属性在应用于噪声流数据时会提高性能,但在传统机器学习数据集上表现较差<sup>[29]</sup>。MUSSELLE 研究了事件流对 DCA 的影响,发现 DCA 对流数据的延迟具有鲁棒性<sup>[30]</sup>。

结合以上研究结果, GREENSMITH 将流作为输入,基于 dDCA 对算法进行函数化的声明和表达,并使用 Haskell 进行实现<sup>[19]</sup>。Haskell 是一种纯函数式编程语言,其中程序的规范被解释为其实现。hDCA 规范由集合上的纯数学函数组成,与之前算法使用

一系列抽象步骤来呈现的方式不同。该规范适用于多数编程语言,也可用于正式推理算法<sup>[19]</sup>。

### 1.3 问题描述

xDCA、dDCA 和 hDCA 在信号提取和信号分类时均采用了主成分分析法,这使得进行危险信号和安全信号选取时依赖于人工经验,一旦确定之后就不可更改,不具备自适应性。而数字微分方法可以自适应地提取信号并随机动态采样抗原,去除信号提取和敏感的数据序列。因此,本文提出 ndhDCA 模型,在输入数据的预处理阶段引入数字微分方法描述数据变化将导致“危险”这一特性,并利用数据的变化自适应提取需要的输入信号。

## 2 ndhDCA 模型

### 2.1 模型框架

ndhDCA 模型架构包括数据采集、数据预处理、信号融合、抗原背景评估与分类以及函数式描述模块,如图 1 所示。

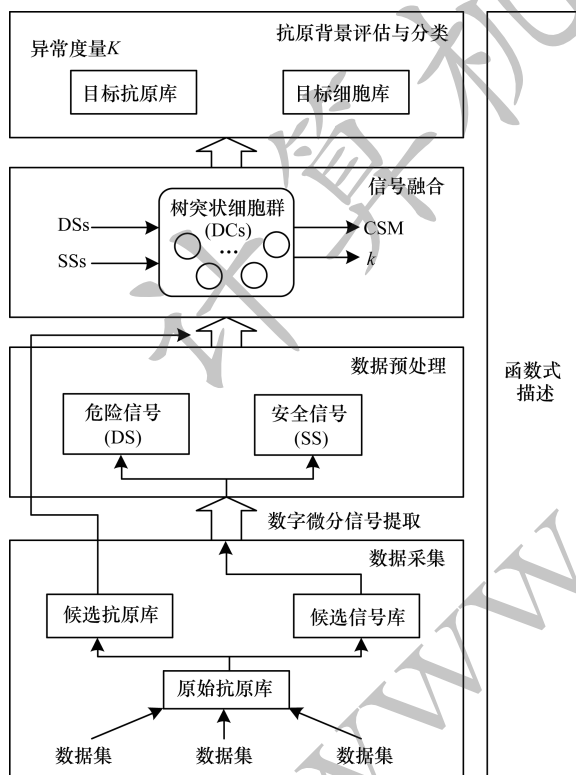


图 1 ndhDCA 模型架构

Fig.1 Framework of ndhDCA model

数据采集模块通过从实际应用环境(或已有的数据集中)获得的所有特征建立原始抗原库,再从中构建候选抗原库和候选信号库,以便后续处理和评估抗原。

数据预处理模块将数字微分用于选取候选信号库中的信号,通过深入的数据分析检测每个时间窗下数据的变化,在考虑特征变化和特征之间相互作

用的前提下,自适应地提取危险信号和安全信号,将其作为信号融合阶段的输入。

信号融合模块利用人工树突状细胞群(DCs)对抗原进行取样,从信号集中获取危险信号和安全信号。当树突状细胞成熟时,通过累积输出信号浓度来确定抗原环境。

抗原环境评估与分类模块通过每次累积的数据计算环境评估的度量值 MCAV 和 K 值,以确定抗原环境处于异常还是正常。

函数式描述模块对上述过程分别进行函数化描述,这也是模型最重要的部分。

### 2.2 基础理论

通常,数据流是一组有时间序列的数据。流是一个有限的列表。任意集合  $X$  上的列表集用  $List_X$  表示,有以下两种情况:

1) 空列表  $\varepsilon$  是  $List_X$  的一个元素。

2) 如果  $x \in X$  且  $xs \in List_X$ , 则  $x:xs \in List_X$ ,  $x:xs$  读作“ $x$  cons  $xs$ ”。

由此可知,空表  $\varepsilon$  是任意集合  $X$  上列表集合的元素,只要  $X$  具有至少一个元素,那么就可以利用上述第 2 种情况将  $X$  的每个元素添加到  $List_X$  中任何列表的开头,因此,  $List_X$  就具有无限数量的元素。例如,  $X = \{1\}$ , 由  $1:\varepsilon \in List_X$  可以推出  $1:(1:\varepsilon) \in List_X$ 。

因为流总是有限的,所以可用列表相同的方式对其进行归纳描述,但需要去掉空列表的情况,以  $Stream_X$  表示任意集合  $X$  上的流集合,其可以被描述为“若  $x \in X, xs \in Stream_X$ , 则  $x \prec xs \in Stream_X$ ”。

### 2.3 数字微分信号提取

在数据预处理阶段,输入由各指标数据构成,数字微分用于从指标集  $\{m_1, m_2, \dots, m_n\}$  中选取特定指标作为危险信号或安全信号。将此阶段的输入信号称为原始数据流,它由各指标和时间戳  $T$  组成,表示为:

$$M = \text{Stream}(T \times m_1 \times m_2 \times \dots \times m_n) \quad (1)$$

用  $s_0$  表示危险信号,安全信号  $s_1$  表示  $s_0$  和  $s_1$  都由原始数据流经过数字微分选取指标并融合各指标后得到,将此过程称为 Mix, 则  $s_0$  和  $s_1$  可以被表示为:

$$s_0, s_1 = \text{Mix}[\text{Stream}(T \times m_1 \times m_2 \times \dots \times m_n)] \quad (2)$$

上述过程主要包括以下 3 个步骤:

1) 将各指标归一化,用数字微分计算每个指标  $m_i$  的变化率。

假设指标与时间之间的映射关系为  $f_{m_i}(t_i)$ , 指标  $m_i$  左右侧可描述为:

$$f_{m_i}(t_i)_{\text{left}} = \frac{(f_{m_i}(t_i) - f_{m_i}(t_{i-1}))}{(t_i - t_{i-1})} \quad (3)$$

$$f_{m_i}(t_i)_{\text{right}} = \frac{(f_{m_i}(t_{i+1}) - f_{m_i}(t_i))}{(t_{i+1} - t_i)} \quad (4)$$

指标  $m_i$  的变化趋势可描述为:

$$f_{m_i}(t_i) = f_{m_i}(t_i)_{\text{right}} - f_{m_i}(t_i)_{\text{left}} \quad (5)$$

上述过程可通过数字微分实现,将其称为 numeric,表示为:

$$\begin{aligned} \text{numeric} : m_i \times T &\rightarrow \Delta m_i \\ \text{numeric}(m_i, T) &= f_{m_i}(t_i)_{\text{right}} - f_{m_i}(t_i)_{\text{left}} \end{aligned} \quad (6)$$

2) 对指标  $m_i$  进行分类,判定是作为危险信号  $s_0$  还是安全信号  $s_1$ ,将此过程称为 divide,设判定的阈值为  $k$ ,则有:

$$\begin{aligned} \text{divide} : m_i &\rightarrow s_i \{ m_1, m_2, \dots, m_n \} \\ \text{divide}(m_i, k) &= \begin{cases} \text{add}_{s_0}(m_i), \text{numeric}(m_i) > k \\ \text{add}_{s_1}(m_i), \text{其他} \end{cases} \end{aligned} \quad (7)$$

3) 累计多个指标融入信号中,作为下一步过程的输入。将累计到信号  $s_i$  的过程命名为  $\text{add}_{s_i}$ ,危险信号  $s_0$  的累积过程如下:

$$\begin{aligned} \text{add}_{s_0} : m_i &\rightarrow s_i \\ \text{add}_{s_0} &= \frac{\sum m + m_i}{n + 1} \end{aligned} \quad (8)$$

## 2.4 信号融合

### 2.4.1 输入

ndhDCA 模型两个输入均为数据流,一个称为事件流(抗原流),另一个称为信号流。事件(抗原)是  $E \times T$  的元素,其中,  $E$  是抗原(事件)的类型集合,  $T$  是时间戳集合。因而事件流可以被定义为:

$$A = \text{Stream}(E \times T) \quad (9)$$

信号流的元素  $S$  是集合  $T \times \mathbb{R}_1 \times \mathbb{R}_2 \times \dots \times \mathbb{R}_n$  的元素,由于算法只有危险信号和安全信号两个输入,它们来自于预处理阶段的结果,因此信号流可以被定义为:

$$S = \text{Stream}(T \times s_0 \times s_1) \quad (10)$$

### 2.4.2 树突状细胞

人工树突状细胞接收输入信号并用他们计算 3 个决策信号,即激活信号、抑制信号和迁移信号。用实数三元组表示这三个信号:

$$\Omega = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \quad (11)$$

每个树突状细胞由事件集、3 个信号值和迁移阈值组成。事件缓冲区  $es$  用于记录细胞在其生命周期中遇到的事件,迁移阈值决定了细胞的生命周期。因此,细胞可被定义为:

$$\text{Cell} = P(E \times T) \times \Omega \times \mathbb{R} \quad (12)$$

给定迁移阈值  $d$ ,则初始化新细胞的方法可以表示为:

$$\begin{aligned} \text{new} : \mathbb{R} &\rightarrow \text{Cell} \\ \text{new}(d) &= (\emptyset, (0.0, 0.0, 0.0), d) \end{aligned} \quad (13)$$

定义以下方法判定细胞是否达到生命周期,如果达到迁移阈值,则评估为 true,否则为 false:

$$\begin{aligned} \text{dead} : \text{Cell} &\rightarrow \text{B} \\ \text{dead}(es, (\omega_A, \omega_I, \omega_M), d) &= \omega_M \geq d \end{aligned} \quad (14)$$

如果某个细胞的  $dead$  值为 true,则对该细胞使用  $reset$  方法重置事件缓冲区和决策信号,并保持原始的迁移阈值:

$$\begin{aligned} \text{reset} : \text{Cell} &\rightarrow \text{Cell} \\ \text{reset}(es, os, d) &= \text{new}(d) \end{aligned} \quad (15)$$

当细胞达到迁移阈值时,计算细胞的临时异常分数,然后将其分配给细胞事件缓冲区的每个事件。有以下 2 种度量方法:

1) 利用成熟背景抗原值 MCAV,返回特定事件被分类为异常的概率,定义如下:

$$\begin{aligned} \text{score}_B : \text{Cell} &\rightarrow \mathbb{R} \\ \text{score}_B(s, (\omega_A, \omega_I, \omega_M), d) &= f(x) = \begin{cases} 1.0, \omega_A > \omega_I \\ 0.0, \text{其他} \end{cases} \end{aligned} \quad (16)$$

2) 如果激活信号  $\omega_A$  大于抑制信号  $\omega_I$ ,则细胞将其缓冲区的事件判定为异常,要计算细胞的实际度量,需要获取  $\omega_A$  减去  $\omega_I$  的值,这是一种真实度量(在 dDCA 中称为  $K_a$ ):

$$\begin{aligned} \text{score}_R : \text{Cell} &\rightarrow \mathbb{R} \\ \text{score}_R(s, (\omega_A, \omega_I, \omega_M), d) &= \omega_A - \omega_I \end{aligned} \quad (17)$$

$\text{score}_R$  计算结果的数字越大,表示细胞事件缓冲区中的事件异常可能性就越大。

### 2.4.3 信号融合

信号融合由以下 3 个步骤组成:

1) 用信号流中当前元素获得的信号值,迭代更新细胞总群中所有细胞的决策信号,此过程称为  $\text{update}_s$ ,其中,  $N$  表示细胞总群,  $p$  表示当前的一个细胞:

$$\begin{aligned} \text{update}_s : (T \times R_1 \times R_2) \times N &\rightarrow N \\ \text{update}_s((_, s_0, s_1), p) &= \\ \{ (es, \text{accumulate}(d, \text{transduction}(s_0, s_1), t)) \mid (es, d, t \in p) \} \end{aligned} \quad (18)$$

更新细胞决策信号阶段分为两个部分,其中,  $\text{transduction}$  把当前输入的信号值映射到细胞各个决策信号,将输入的危险信号  $s_0$ 、安全信号  $s_1$  转化为决策信号。 $\text{accumulate}$  将当前的决策信号累加到先前决策信号,计算出细胞新的决策信号。为计算决策信号,本文应用线性函数,用  $\omega = \omega_j \times s_i + \dots + \omega_j \times s_i$  形式的公式来表示,其中,  $\omega$  是正在计算的判定信号,  $s_i$  是输入信号,  $\omega_j$  是权重。权重可从生物免疫学对树突状细胞的实验中获得,如表 1 所示。

表 1 信号融合权重  
Table 1 Fused weights of signals

信号	$\omega_A$	$\omega_I$	$\omega_M$
$s_0$	1.0	0.0	1.0
$s_1$	-2.0	1.0	1.0

由表 1 中的权重得出决策信号计算公式:

$$\omega_A = s_0 + (-2.0) \times s_1$$

$$\omega_I = s_1$$

$$\omega_M = s_0 + s_1 \quad (19)$$

由此,transduction 的函数表达式可定义为:

$$\text{transduction}(s_0, s_1) =$$

$$(s_0 + (-2.0) \times s_1, s_1, s_0 + s_1) \quad (20)$$

一旦计算了来自两个输入信号的决策信号的值,则需要将它们添加到细胞的决策信号的当前值,由此 accumulate 可被描述为:

$$\text{accumulate}((\omega_A, \omega_I, \omega_M), (\omega'_A, \omega'_I, \omega'_M)) =$$

$$(\omega_A + \omega'_A, \omega_I + \omega'_I, \omega_M + \omega'_M) \quad (21)$$

2)更新决策信号导致每个细胞的迁移值增加,即  $\omega_M$  增加。因此,需要检查细胞是否已超过相应迁移阈值。当寿命超过其迁移阈值的细胞,需要使用 results 函数为其事件缓冲区中的每个事件生成中间异常分数。

$$\text{results}(p) =$$

$$\{(e, \text{score}(c) | c \in p, \text{dead}(c), e \in \text{events}(c))\} \quad (22)$$

3)达到迁移阈值的细胞内异常分数值计算完成后,使用 migrate 重置这些细胞以生成新一代:

$$\text{migrate}(cs) =$$

$$\{\text{if } \text{dead}(c) \text{ then } \text{reset}(c) \text{ else } c | c \in cs\} \quad (23)$$

migrate 遍历群体中所有细胞,并用 dead 测试是否超过其迁移阈值。如果 dead 为 true,则通过 reset 重置细胞,清除细胞的事件缓冲区并将其决策信号设置为 0,迁移阈值保持不变。

## 2.5 抗原背景环境评估与分类

在经过  $n$  次迭代之后,停止处理输入并返回映射到异常分数的一组事件的元素。该集合被传递给 analyze 函数。此时,对于同一事件类型,可能会有多个临时异常分数,在 analyze 中计算每种事件类型的平均异常分数:

$$\text{analyse}(\Phi) =$$

$$\{(e, \frac{\sum_v v}{|v|} | e \in E)\} \text{ where } v = \{s | ((e', t), s) \in \Phi, e = e'\} \quad (24)$$

## 3 实验与分析

### 3.1 实验设计

#### 3.1.1 数据集选择

本文提出 ndhDCA 的目的是在信号提取阶段能够脱离人工经验,自适应地提取 DS 和 SS 信号,同时保证分类的准确性。为展示 ndhDCA 的有效性及其性能,测试本文模型的性能,将使用 ndhDCA、DCA、dDCA/hDCA 分别在两个数据集进行对比实验。

实验在两个数据集上进行:

1) the UCI Wisconsin Breast Cancer dataset(WBC), 其由 700 个数据实例组成,每个数据项有 10 个特征。

2) KDD99 数据集,其源自 DAPRA 98 Lincoln Lab 数据集,用于将数据挖掘技术应用于入侵检测领域,由约 500 万个数据实例组成,每个数据实例具有 42 个特征。本文使用是其 10% 的子集,数据项从整个数据集中随机按比例选择。

在两个数据集的数据项中,WBC 数据集按种类排序,为有序数据;KDD99 数据集经由多次随机打乱,为无序数据。

#### 3.1.2 参数说明

数据集中每个数据项被映射为抗原。在所有实验中,DC 总数均设置为 100,每个循环中对 10 个 DC 采样,以确保 ndhDCA 的自适应性。在数字微分计算结果中,选取变化最大的 4 个信号形成 DS,最小的一个信号为 SS,其余所有特征参数设置按照文献[1]设置,这些参数来自经验免疫学数据(同 hDCA/dDCA)。对于 DCA,在 WBC 和 KDD99 中设置的异常值分别为 0.65 和 0.446。

### 3.2 实验结果与分析

为与原始 DCA 进行对比,本文使用的编码环境为 Python 3.6。实际上,本文给出的函数化表达式更适合在 Haskell 的环境下进行实验。

#### 3.2.1 评价指标说明

本文实验使用分类器模型常用指标作为评价指标,使用的每个指标的具体说明如表 2 所示。

表 2 评价指标说明

Table 2 Description of evaluation indexes

评价指标	指标解释
阳性预测率 (PPV)	被分类为阳性的样本中分类正确的占比
阴性预测率 (NPV)	被分类为阴性的样本中分类正确的占比
召回率 (RN)	所有阳性样本中被分类正确的占比
特异度 (S)	所有阴性样本中被分类正确的占比
平均值 (AVG)	PPV、NPV、RN、S 指标的平均值
漏报率 (FAR)	被分类为阴性的样本中分类错误的占比
马修斯相关系数 (MCC)	描述实际分类与预测分类之间的相关系数,取值范围为 $[-1, 1]$ , 0 为随机预测,越接近于 1 说明分类能力越强
AUC 值	ROC 曲线下面积,直观反映分类能力, AUC 为 0.5 时为随机分类器,越接近 1 说明分类能力越强
准确率 (PRE)	预测准确率

#### 3.2.2 在有序数据集 WBC 上的实验

有序数据集 WBC 上 3 种算法的对比实验结果如表 3 和图 2 所示。从实验结果看,ndhDCA 在阳性预测率、阴性预测率、特异度上都有更高的评价,召回率与 dDCA 相同,马修斯相关系数、AUC 高于其他两种算法。漏报率也低于 dDCA 和 DCA。从 ROC 曲线图可以更明确地看出,本文提出的 ndhDCA 不仅能自适应地提取危险信号和安全信号,其在有序数据集上,与 DCA、hDCA/dDCA 相比,还具有更高的准确率和更低的误报率。

表 3 WBC 数据集上的实验结果

Table 3 Experimental results on the WBC dataset

%

模型	PPV	NPV	RN	S	FAR	MCC	AUC	AVG	PRE
DCA	19.83	95.25	90.83	95.87	4.75	86.97	66.500	93.49	94.14
dDCA	86.08	98.83	97.92	91.74	1.17	87.25	94.827	93.64	93.86
ndhDCA	94.38	98.89	97.92	96.96	1.11	94.07	97.440	97.04	97.29

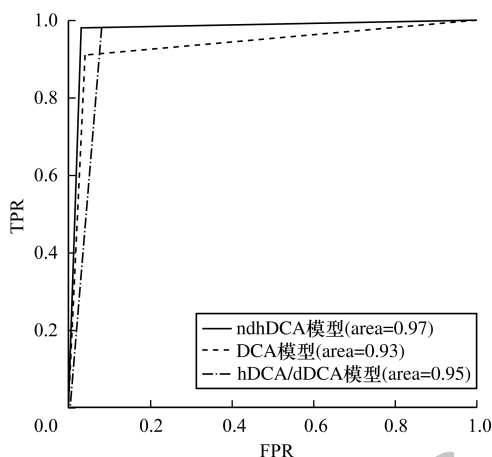


图 2 WBC 数据集上的 ROC 图

Fig.2 ROC graph on the WBC dataset

表 4 KDD99 数据集上的实验结果

Table 4 Experimental results on the KDD99 dataset

%

模型	PPV	NPV	RN	S	FAR	MCC	AUC	AVG	PRE
DCA	0.00	55.40	0.00	100.00	44.60	0.00	50.00	38.85	55.40
dDCA	45.36	72.64	97.40	5.56	27.36	7.30	51.48	51.48	46.52
ndhDCA	55.35	78.90	85.15	44.69	21.10	31.98	64.93	66.03	62.74

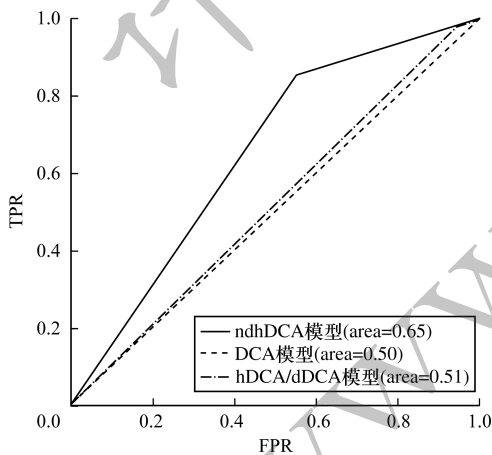


图 3 KDD99 数据集上的 ROC 图

Fig.3 ROC graphs on the KDD99 dataset

#### 4 结束语

本文构建了基于数字微分理论的函数化树突状细胞模型 ndhDCA。在 hDCA 模型基础上,利用数字微分可根据离散数值变化和变化趋势感知危险的特性,以 DC 信号随机动态地对抗原进行采样,通过分析数据变化及变化趋势确定并自适应提取 DS 和 SS 信号。实验结果表明,相比通过人工经验设置,

#### 3.2.3 在无序数据集 KDD99 上的实验

无序数据集 KDD99 上 3 种算法的对比实验结果如表 4 和图 3 所示。在之前的 DCA 研究中,错误的分类常发生在过渡边界。因此,当环境连续多次变化时 DCA 会产生更多错误。ndhDCA 每次运行对每个数据项进行 10 次采样,可以缓解数据项混乱带来的影响。

实验结果显示,由于数据随机乱序了多次,DCA 的性能基本上等同于随机分类器,对于 dDCA 和 ndDCA,在阳性预测率、阴性预测率、特异度、马修斯相关系数、准确率上 ndhDCA 有更高的评价,但误报率比 dDCA 高。图 3 所示的 ROC 图更清晰地反映了在无序数据集 KDD99 上,3 种算法分类效果均不理想,但在数据项乱序的情况下,ndhDCA 相比较于其他两种算法取得了更准确的分类结果。

该模型的算法普适性更强,并能减小输入数据顺序的敏感性。后续将评估本文模型在无序数据集上的分类效果,并进一步提高其检测性能。

#### 参考文献

- [1] CASTRO L R D, TIMMIS J. Artificial immune systems; a new computational intelligence paradigm [M]. Berlin, Germany: Springer, 2002.
- [2] LU Tianliang, ZHANG Lu, CAI Manchun, et al. shellcode detection algorithm inspired by hyper-ellipsoids immune theory [J]. Journal of Chinese Computer Systems, 2018, 39(6): 1255-1259. (in Chinese)
- [3] 芦天亮, 张璐, 蔡满春, 等. 一种超椭圆免疫理论启发的 shellcode 检测算法 [J]. 小型微型计算机系统, 2018, 39(6): 1255-1259.
- [4] DUDEK G. Artificial immune system with local feature selection for short-term load forecasting [J]. IEEE Transactions on Evolutionary Computation, 2017, 21(1): 116-130.
- [5] GREENSMITH J, AICKELIN U. Dendritic cells for SYN scan detection [C] // Proceedings of Conference on Genetic and Evolutionary Computation. New York, USA: ACM Press, 2007: 49-56.
- [6] AL-HAMMADI Y, AICKELIN U, GREENSMITH J. DCA for bot detection [C] // Proceedings of IEEE World Congress on Computational Intelligence. Washington D. C., USA: IEEE Press, 2008: 1-10.

- [6] OATES R, GREENSMITH J, AICKELIN U, et al. The application of a dendritic cell algorithm to a robotic classifier[C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2007:204-215.
- [7] SILVA G C, CAMINHAS W M, ERRICO L D. Dendritic cell algorithm applied to ping scan investigation revisited: detection quality and performance analysis[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2017, 1(4): 236-247.
- [8] IGBE O, AJAYI O, SAADAWI T. Detecting denial of service attacks using a combination of dendritic cell algorithm and the negative selection algorithm[C]//Proceedings of IEEE International Conference on Smart Cloud. Washington D. C., USA: IEEE Press, 2017:1-6.
- [9] IGBE O, DARWISH I, SAADAWI T. Deterministic dendritic cell algorithm application to smart grid cyber-attack detection[C]//Proceedings of IEEE International Conference on Cyber Security and Cloud Computing. Washington D. C., USA: IEEE Press, 2017:1-5.
- [10] GAN Ying, LIANG Yiwen, TAN Chengyu, et al. Earthquake prediction method based on danger theory[J]. Computer Engineering, 2019, 45(1): 278-283. (in Chinese)  
甘颖, 梁意文, 谭成予, 等. 基于危险理论的地震预测方法[J]. 计算机工程, 2019, 45(1): 278-283.
- [11] YANG Chao, QIN Tingdong, FAN Bo, et al. Study on detection of Weibo spammers based on danger theory in artificial immunity system[J]. Computer Science, 2018, 45(11): 145-149, 166. (in Chinese)  
杨超, 秦廷栋, 范波, 等. 基于人工免疫危险理论的微博水军用户检测研究[J]. 计算机科学, 2018, 45(11): 145-149, 166.
- [12] GREENSMITH J, AICKELIN U, CAYZER S. Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection[C]//Proceedings of International Conference on Artificial Immune Systems. New York, USA: ACM Press, 2005:153-167.
- [13] GREENSMITH J, AICKELIN U, TWYXCROSS J. Articulation and clarification of the dendritic cell algorithm[C]//Proceedings of International Conference on Artificial Immune Systems. New York, USA: ACM Press 2006:404-407.
- [14] GU F, GREENSMITH J, AICKELIN U. The dendritic cell algorithm for intrusion detection[EB/OL]. [2019-05-10]. <https://arxiv.org/ftp/arxiv/papers/1305/1305.7416.pdf>.
- [15] GREENSMITH J, AICKELIN U. The deterministic dendritic cell algorithm[C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2008:291-302.
- [16] YANG Chen, LIANG Yiwen, TAN Chengyu, et al. Optimized dendritic cell algorithm combined with XGBoost[J]. Computer Engineering, 2019, 45(9): 194-197, 203. (in Chinese)  
杨晨, 梁意文, 谭成予, 等. 结合 XGBoost 改进的树突状细胞算法[J]. 计算机工程, 2019, 45(9): 194-197, 203.
- [17] WANG Shuyang, MU Xiaodong, ZHANG Li. Improved dendritic cell algorithm integrated with principal component analysis[J]. Computer Engineering and Design, 2017, 38(6): 1414-1417. (in Chinese)  
王舒洋, 慕晓冬, 张力. 集成 PCA 的改进树突状细胞算法[J]. 计算机工程与设计, 2017, 38(6): 1414-1417.
- [18] GU F, GREENSMITH J, AICKELIN U. Theoretical formulation and analysis of the deterministic dendritic cell algorithm[J]. Biosystems, 2013, 111(2): 127-135.
- [19] GREENSMITH J, GALE M B. The functional dendritic cell algorithm: a formal specification with Haskell[C]//Proceedings of IEEE Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2017:1787-1794.
- [20] CHELLY Z, ELOUEDI Z. A new data pre-processing approach for the dendritic cell algorithm based on fuzzy rough set theory[C]//Proceedings of GECCO'13. New York, USA: ACM Press, 2013:163-164.
- [21] LIANG Yiwen, CAO Linglin, CAI Ying. Introduction to danger sensed through numerical differential[J]. Journal of Harbin Engineering University, 2006, 27(s1): 228-232. (in Chinese)  
梁意文, 曹玲林, 蔡瀛. 危险感知的数字微分初步[J]. 哈尔滨工程大学学报, 2006, 27(s1): 228-232.
- [22] YANG Chao, LI Tao. Research of danger signal extraction based on changes in danger theory[J]. Computer Science, 2015, 42(8): 170-174. (in Chinese)  
杨超, 李涛. 计算机免疫危险理论中危险信号的提取研究[J]. 计算机科学, 2015, 42(8): 170-174.
- [23] ZHOU Wen, LIANG Yiwen, DONG Hongbin, et al. A numerical differentiation based dendritic cell model[C]//Proceedings of IEEE International Conference on Tools with Artificial Intelligence. Washington D. C., USA: IEEE Press, 2017:1-5.
- [24] XIAO Zhenhua, LIANG Yiwen, TAN Chengyu, et al. Dendritic cell fault detection method based on numerical differentiation[J]. Acta Electronica Sinica, 2019, 47(5): 1029-1035. (in Chinese)  
肖振华, 梁意文, 谭成予, 等. 基于数值微分的树突状细胞故障检测方法[J]. 电子学报, 2019, 47(5): 1029-1035.
- [25] CHELLY Z, ELOUEDI Z. FDCM: a fuzzy dendritic cell method[C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2010: 102-115.
- [26] GU Feng. Theoretical and empirical extensions of the dendritic cell algorithm[D]. Nottingham, UK: University of Nottingham, 2011.
- [27] OATES R, KENDALL G, GARIBALDI J. Frequency analysis for dendritic cell population tuning: decimating the dendritic cell[J]. Evolutionary Intelligence, 2008, 1: 145-157.
- [28] STIBOR T, OATES R, KENDALL G, et al. Geometrical insights into the dendritic cell algorithm[C]//Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation. New York, USA: ACM Press, 2009:1275-1282.
- [29] GU F, FEYEREISL J, OATES R, et al. Quiet in class: classification, noise and the dendritic cell algorithm[C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2010:173-186.
- [30] MUSSELLE C J. Insights into the antigen sampling component of the dendritic cell algorithm[C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2010:88-101.