文章编号:1007-6735(2012)06-0541-04

基于树突状细胞算法的入侵检测系统研究

钟姗姗, 周健勇
(上海理工大学 管理学院,上海 200093)

摘要:对树突状细胞算法在入侵检测系统中的应用进行研究.对树突状细胞算法中抗体集合的确定和阈值设置进行动态化改进,以建立动态树突状细胞算法.同时,采用UCI数据集进行对比实验.将动态树突状细胞算法与传统树突状细胞算法的检测结果进行比较,结果表明,动态树突状细胞算法提高了处理数据的有效性、鲁棒性和自适应性.

关键词:危险理论;树突状细胞算法;入侵检测
中图分类号:TP 391 **文献标志码**:A

Intrusion Detection System Based on Dendritic Cells Algorithm

ZHONG Shan-shan, ZHOU Jian-yong
(*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: The dynamic dendritic cells algorithm in the intrusion detection system was introduced. The algorithm possesses the capability of dynamic determination of antibodies set and can improve the threshold setting. The comparative experiment on detection results between traditional dendritic cell detection algorithm and dynamic dendritic cell algorithm based on UCI datasets illustrates that the dynamic dendritic cell algorithm effectively improves the accuracy of the data processing and increases the robustness and self-adaptability.

Key words: *danger theory*; *dendritic cells algorithm*; *intrusion detection*

随着网络信息技术的发展,网络应用中潜在的信息安全问题也在不断增多.各类针对系统的入侵行为的破坏程度不容忽视,其中程度严重的破坏可造成相当于战争行为的危害.在1980年Anderson[1]将入侵行为定义为是具有潜在可能性,在未经授权情况之下,企图进行如下操作:

a. 访问信息;

b. 操作信息;

c. 导致系统不可靠或无法使用.

为防范此类诸多的入侵行为,不同的防范措施应运而生.根据有关统计资料表明,30%的入侵行为发生在有防火墙的情况之下,因而作为防火墙的合理补充的入侵检测系统应运而生.受到生物系统启发的人工免疫原理被证明对防范入侵行为具有高效性、自组织性、高适应性和鲁棒性,符合入侵检测的需求.本文从树突状细胞功能原理中受到启发,通过

收稿日期:2012-03-22
作者简介:钟姗姗(1988-),女,硕士研究生.研究方向:信息安全.E-mail:zssd586@yahoo.cn
周健勇(1970-),男,副教授.研究方向:系统优化、系统工程和信息安全.E-mail:zjy@mega-tel.net

应用人工免疫原理中的危险理论和树突状细胞算法对入侵检测方法进行研究.

1 树突状细胞算法

1.1 人工免疫和危险理论概述

免疫系统是一个由免疫活性分子、免疫细胞、免疫组织和器官组成,分布于人体各个部分的复杂、自适应系统.自1891年Ehrlich建立免疫学概念以来,对于免疫原理的研究主要集中于自我非自我(self-non-self,SNS)免疫学古典学说,即通过细胞选择来区分"自体"与"非自体"蛋白分子.但这一学说对有些问题仍然无法解释,比如,在免疫系统只对非自体作出应答情况下,为何会出现自身免疫性疾病,以及为何在肠道环境中数以百万计的细菌并没有触发免疫应答.对于这些SNS学说无法解释的问题,免疫学家Matzinge[2]给出了解答,在1994年他提出免疫系统的本质并非在于区分"自体"与"非自体",而是根据细胞是否受到损伤的信号来作出反应.这一理论即被称为危险理论,该理论表明免疫系统是依据细胞异常死亡而产生的危险信号来作出应答.

危险理论(danger theory,DT)[2]包含对免疫应答的激活和抑制,即在机体组织内出现的危险信号足以激活免疫系统,而另一类非危险信号可以防止免疫反应.这种抑制机制是由细胞凋亡引起的结果,是细胞从人体中迁移出去的正常行为.在一个细胞发生这种细胞凋亡的过程中,它会向周围环境释放出各种信号.树突状细胞(DC)也对这类信号的浓度变化异常敏感,并能够对危险和安全信号进行信息融合,以确定机体组织环境内遇险还是工作正常.危险理论认为,免疫系统只在危险出现时进行响应或者处于积极抑制状态.

在危险理论中相关信号是由树突状细胞进行提呈.树突状细胞作为最强专职抗原提呈细胞(APC),负责对机体组织内的病原体相关分子模式(PAMP)进行收集、识别、分析与处理,最后提呈给免疫细胞.显见,树突状细胞对整个诱导特异性免疫应答过程起到了至关重要的作用.

1.2 传统树突状细胞算法

树突状细胞算法(dendritic cells algorithm,DCA)[3-4]主要是针对DC细胞运作过程的仿生原理而设计,其单个DC决策作用如图1所示.该算法通过信号与抗原入侵的双重刺激,确认入侵行为.在细胞受损或细菌入侵产生病原相关分子模式之后,未成熟DC(iDC)开始采集抗原群和信号集,即安全信号(SS)、危险信号(DS)、病原相关分子模式.T_h 为阀值,k 为成熟信号.

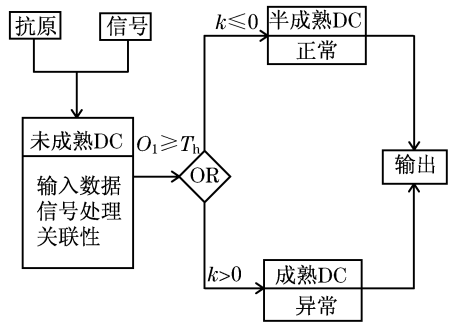

图1 DCA流程

Fig.1 DCA flow chart

将表1中的权值代入式(1)[5-7],计算得到输出信号.

$$O_j = \sum_i^3 W_{ij} S_i \Big/ \sum_i^3 W_{ij} \tag{1}$$

式中,O_j 为输出信号CSM、SemiDC、finDC的输出值;W_{1j} 为对应于输出信号 j 的输入信号PAMPs的权值;W_{2j} 为对应于输出信号 j 的输入信号DS的权值;W_{3j} 为对应于输出信号 j 的输入信号SS的权值;S_1 为输入信号PAMPs的值;S_2 为输入信号DS的值;S_3 为输入信号SS的值.

表1 DCA权值表

Tab.1 Weights used for signal processing

输入信号	相对输出信号权重 W_{ij}		
	O_1	O_2	O_3
S_1	2	0	2
S_2	1	0	1
S_3	2	2	−2

表1是经多次生物实验后所得出的输入信号与输出信号之间的具体权值.根据式(1)计算得出 O_1,即协同刺激信号(CSM)值,当 O_1 大于阈值 T_h 时,则发生状态转移;反之,则重新开始采集抗原群与信号集.如图1所示,对迁移后的抗原根据式(1)计算成熟信号 k.若DC释放的成熟状态信号 $k>0$,则提呈抗原为半成熟DC(semiDC)转成熟状态;反之,则提呈为成熟DC(finDC)转半成熟状态.

2 动态树突状细胞算法

2.1 抗体集的确定

同种抗体浓度[8]是受到控制的,浓度高的抗体

系统要对其进行抑制，同样，浓度低的抗体则要繁殖以保持多样性. 记忆抗体集与抗原的分布关系有两种极端情况：

$$\begin{cases} Ab_1 \cup Ab_2 \cup \cdots \cup Ab_i = Ag \\ Ab_1 \cap Ab_2 \cap \cdots \cap Ab_i = \Phi \end{cases}$$

$$\begin{cases} Ab_1 \cup Ab_2 \cup \cdots \cup Ab_i = Ab_j \\ Ab_1 \cap Ab_2 \cap \cdots \cap Ab_i = Ab_j \end{cases}$$

式中，Φ 为空集；Ab_i，Ab_j 代表抗体子集；Ag 为抗原全集，即抗体子集的并为抗原全集，交为空集；或者抗体子集的并为该抗体子集 j，且交为该抗体子集 j. 前者高度独立，当任一抗体子集失效，其它子集将不能有效检测到失效抗体子集所覆盖区域，这将给系统带来极大危险性. 而后者抗体重叠，存在大量冗余和不必要的资源浪费. 所以，抗体子集间存在适度的交叉，能够很好地提高系统多样性，并防止鲁棒性退化. 对抗体浓度进行度量，旨在保持抗体多样性的同时，不至于收敛到某一区域.

在非空系统集合 S 上，通过计算抗体 Ab_i 的 Euclid 距离，判断抗体间是否相似，其基于距离的浓度函数为

$$C(Ab_i) = 1 - \frac{\sum_{j=1, i\neq j}^{n} \| Ab_i - Ab_j \|}{\sum_{i=1, j=1, i\neq j}^{n} \| Ab_i - Ab_j \|}$$

式中，$C(Ab_i)$为抗原在抗原集中的浓度.

抗体亲和度[8]表征抗原和抗体亲和力的拟合度，高亲和度代表抗体和抗原发生亲和作用的可能性高. 抗原和抗体的亲和度函数为

$$D(Ab_i) = \frac{1}{1 + f'(Ab_i)}$$

$$f'(Ab_i) = \frac{f_{\max} - f(Ab_i)}{f_{\max} - f_{\min}}$$

式中，$D(Ab_i)$为抗体与抗原间的亲和度；$f(Ab_i)$为抗体 i 的适应度值；$f_{\max}$，$f_{\min}$为适应度最大值和最小值.

理想的抗体集应同时具有高亲和度、低浓度，即

$$E(Ab_i)_{\max} = \frac{D(Ab_i)}{C(Ab_i)}$$

式中，$E(Ab_i)_{\max}$为抗体与抗原间亲和度同抗原在抗原集中浓度的最大比值.

2.2　动态 DCA 算法

采用不同的数据集分割方式进行检测，会给检测结果带来很大差异. 动态 DCA 算法（DDCA）有别于现在普遍使用的基于时间分割数据集的检测算法，而采用基于抗原分割数据集的方式. 基于抗原分割来进行检测能确保对数据集的分析具有更好的解释性，所以，DDCA 采用基于抗原分割，该算法流程如图 2 所示.

为了创建 DDCA 算法，有必要对初始抗原进行定义，计算其亲和度与浓度比值以选定抗体集. 首先，算法仍然需要输入信号，以及抗原的提呈过程. 两类输入信号分别为抑制信号和激活信号，即 SS 信号和 DS 信号. 抗原提呈以动态设定的阈值和计算所得的输出信号值 O_1 为提呈标准. 树突状细胞群内的每个树突状细胞具有相同抗原的输入信号，用同一方式处理这些信号. 在整个树突状细胞群中的输出信号仅计算一次. 通过计算抗原亲和度与浓度比值动态选定抗体集，确定阈值，所以，对信号处理结果有稳定、优化作用.

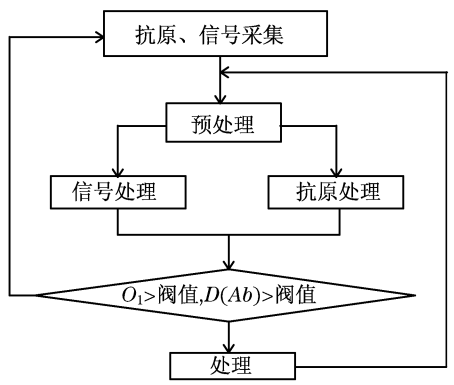

图 2　DDCA 流程

Fig.2　DDCA flow chart

DDCA 的主要框架思想是以 $E(Ab_i)_{\max}$为衡量标准来动态确定抗体集合，$E(Ab_i)_{\max}$可以有效避免抗体集的收敛，同时保证抗体集合的多样性. 对于原 DCA 中的阈值参数，则动态地采用抗体集合中最小的 N 个 CSM 的均值和亲和度值作为阈值参数. 动态设定阈值参数使抗体能更好地适应抗原的多样性，具有强鲁棒性.

3　仿真实验

仿真实验选用标准的美国威斯康星大学医学院的乳腺癌数据集（UCI）[9]作为实验数据，UCI 数据集包含 699 条数据，包含 Class 类在内的 10 个属性，其中，458 条属于良性数据集，241 条属于恶性数据集. 如图 3 所示（见下页），采用 DDCA 的检测结果比朴素贝叶斯算法（Nbayes）和 K 平均算法（K-Means）的结果的准确率高，错警率低，说明该算法具有有效性.

将所有数据分割为上、中、下3个部分，每个部分各含有233条数据.分别以3个不同部分作为训练集，进行模拟实验.每1条数据视为1个乳腺癌细胞，根据9个属性分布情况，选取[cell size]、[cell shape]、[bare nuclei]、[bland chromatin]、[bland chromatin]、[normal nucleoli]各个属性所对应的良性数据集均值与抗原属性值之间的绝对偏差作为危险信号，[clump thickness]偏离良性数据集均值作为安全信号、病原体相关分子模式，结果如表2所示.

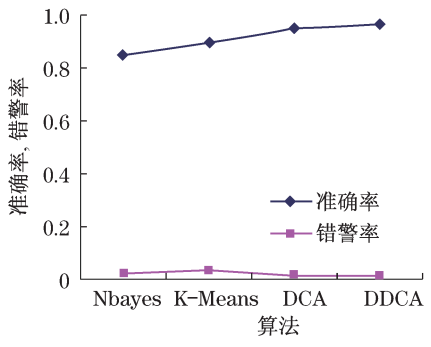

图3 Nbayes、K-Means、DCA、DDCA准确率、错警率

Fig.3 Detection rates and FP rates of Nbayes, K-Means, DCA and DDCA

表2 DDCA检测结果

Tab.2 Detection results of DDCA

数据集	准确率/%	错警率/%
上部分	96.794	1.462
中部分	96.994	1.169
下部分	96.593	2.046

由表2可知，DDCA在采用不同数据段作为训练集时，数据准确率保持在96.5%以上，而错警率也较低，在1.5%左右，显见检测结果稳定，算法具有鲁棒性.

4 结束语

讨论了树突状细胞性能和现有免疫学危险理论在入侵检测中的应用，并提出了一种动态DCA算法，且通过实验进行了验证.在算法中，通过对树突状细胞运作进行模拟，动态地确定危险抗原进行警报.实验结果表明，算法实现了预期的效果，具有高效性、鲁棒性.但模型中仍然存在许多问题，在以后的研究中将继续进行探索.

参考文献：

[1] Anderson J P. Computer security threat monitoring and surveillance [R]. Fort Washington: James P Anderson Company, 1980.

[2] 郭晨，梁家荣，夏洁武.基于危险理论的人工免疫原理与应用[J].计算机应用研究，2007，24(6)：18－21.

[3] Aickelin U, Greensmith J. Sensing danger: innate immunology for intrusion detection [J]. Information Security Technical Report, 2007, 12(4): 218－227.

[4] Greensmith J, Aickelin U, Tedesco G. Information fusion for anomaly detection with the dendritic cell algorithm[J]. Information Fusion, 2010, 11(1): 21－34.

[5] 罗超，郭晨，梁家荣.确定性树突状细胞算法的异常检测系统[J].江西师范大学学报，2011，35(2)：170－172.

[6] 楚赟，戴英侠，万国龙.一个基于免疫的分布式入侵检测系统模型[J].计算机应用，2005，25(5)：1153－1157.

[7] 杨向荣，沈钧毅，罗浩.人工免疫原理在网络入侵检测中的应用[J].计算机工程，2003，29(6)：27－29.

[8] 刘韬.人工免疫系统及其数据挖掘应用研究[M].徐州：中国矿业大学出版社，2010.

[9] 李光，张凤斌.基于树突状细胞算法的分类方法研究[J].电脑知识与技术，2010，6(31)：8798－8800.