基于云模型的识别技术

王思鹏
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430081)

摘 要:以机体免疫思想的人工免疫系统作为一种动态自适应的方法,可以更好地解决传统的网络防御方法的被动、静态等缺点。引入云模型的概念,提出了人工免疫系统中云识别的理论,即根据环境的需要用多个识别器对不确定性抗原进行联合云识别,以达到降低伪肯定率和伪否定率的目的。

关键词:人工免疫系统;云模型;不确定性
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2006)06-0187-04

Detecting Technology Based on Cloud Model

WANG Si-peng
(College of Computer Sci. & Techn., Wuhan University of Sci. and Techn., Wuhan 430081, China)

Abstract: Artificial immune system(AIS) with the feature of adaptability is capable of dealing with dynamic complex changes of external environment, ensuring security of computer system and network. Traditional defending methods have some recognitions such as regarding network intrusion behavior as defending statically, passively and etc. This paper introduces the theory of cloud model and puts forward the theory of cloud detecting which uses multi-detectors to evaluate the uncertain antigen, and will decrease the false positive rate and false negative rate.

Key words: artificial immune system; cloud model; uncertainty

0 引 言

目前,计算机免疫系统主要是解决计算机安全系统中发生频繁的、形式变化多样的入侵和攻击的识别或检测问题,然后可以采取相应的措施进行防御。这种防御能力具有更高的适应性、一般性,通过其自学习、自适应能力来改变传统计算机安全系统中的不断地"发现漏洞→打补丁"的被动防御方法,实现更一般目标的防御,从而扩充计算机安全系统[1,2]。

文中分析了人工免疫学中的不确定性以及这种不确定性在人工免疫系统的建立过程中导致的识别问题:由于抗原的属性随环境的不同具有不确定性[3]。利用云模型的数字特征和相关的云发生器算法,构建了云识别的算法并进行了相关实验,提出使用多个识别器联合对一个外来抗原进行云识别的思想,最后达到降低传统识别方法伪肯定率和伪否定率的目的。

1 人工免疫学中的不确定性及其引发的问题

1.1 人工免疫学中的不确定性

在人工免疫系统中,抗原是海量的、无穷的,不可能穷举,自然也不可能对其一一定性;同时各子系统为了防止单个子系统的脆弱性在整个系统中蔓延而保持子系统之间的多样性。这样共同导致了抗原在环境不同时具有不确定性。比如对无穷抗原空间中的某一个不能定性的抗原 x,由于子系统的不同而对它有不同的定义,安全级别高的定义其为 NONSELF,而安全级别低的定义它为 SELF。这样就导致了对该抗原的判定具有随机(随系统的不同而定)、模糊性(没有精确统一的定义),这两点共同构成了对抗原判定的不确定性。

表1 抗原的不确定性

子系统 / 抗原	子系统一(安全级别高)	子系统二(安全级别低)
抗原一(属于已知 SELF 集)	SELF	SELF
抗原二(不确定性抗原)	NONSELF	SELF
抗原三(被识别器识别)	NONSELF	NONSELF

如表1所示,子系统一和子系统二的安全级别不同,具有多样性。现在有3个外来抗原抗原一、抗原二和抗原三进入系统,人工免疫系统开始识别它们。如表所示,抗原一属于已知的 SELF 集合,因此,子系统一和子系统二都把它当成 SELF。抗原三能够被成熟识别器集里的某个识别器所匹配(文中所指的匹配方式是指 R -邻域位匹配),因此子系统一和子系统二都把它识别为 NONSELF。而抗原二不属于已知的 SELF 集合,而且也不能

收稿日期:2005-11-07
作者简介:王思鹏(1978-),男,湖北武汉人,硕士,助教,研究方向为网络安全、软件工程。

被成熟识别器集里的所有识别器所匹配识别,所以它根据系统的环境不同而具有不确定性,在安全级别高的子系统一中可能为NONSELF,在安全级别低的子系统二中可能为SELF。因此,可以看出在人工免疫系统中随环境的不同抗原的属性具有不确定性。实际防御中,这种不确定性会给人工免疫系统带来一些问题[4]。

1.2 识别过程中的问题

由于抗原空间巨大并且抗原的属性具有不确定性,用识别器逐一单独地去识别一类抗原、并将识别器不能识别也不属于已知SELF集的不确定性抗原全部判定为SELF或NONSELF会导致高的伪肯定率和伪否定率[5]。如何应对这种抗原的不确定性问题是人工免疫学中的一大难题。问题分析如下:

设抗原为 x ,已知的SELF集合为SELF',成熟的识别器集合为 D 。

(1)传统的识别过程如下:如果 x 能被 D 集合中的某个个体识别,那么可以判断它是NONSELF。如果 x 属于SELF',那么可以判断它是SELF。最后如果 x 既不能被 D 中的所有个体识别,也不属于SELF',即属于空白空间,那么 x 的属性具有不确定性。安全级别高的系统把 x 判定为NONSELF,安全级别低的系统把 x 判定为SELF。

(2)在实际应用中,SELF集合往往是一个异常庞大的集合,而搜集到的已知样本SELF'只是SELF集的一个子集,甚至是一个很小的子集。一个抗原不属于已知的SELF集合并不能代表它不是SELF,同理,它不能被所有成熟的识别器识别不能代表它不是NONSELF。因此对于那些既不能被 D 中所有识别器识别也不属于SELF'的不确定的抗原 x 来说,由于子系统的多样性,统一把它们划分为SELF或NONSELF会导致高的误判(伪肯定)和漏判(伪否定)。

对此不确定性导致的问题,解决思路有两个:

a.消除不确定性。由于不确定性是由人工免疫学的某些必不可少的特性比如多样性、适应性等所决定的,因此不能消除不确定性。

b.改变对不确定性抗原的判定方法。寻求一种能较好地处理不确定性问题的工具,并将其引入人工免疫系统的识别过程,改变对不确定性抗原的判定方法,以降低伪肯定率和伪否定率。

文中引入云模型的理论来处理这个不确定性的问题。采用多个识别器对不确定性抗原 x 进行联合决策,从而降低对 x 识别的伪肯定率和伪否定率。

2 不确定性的研究工具——云模型

2.1 云模型的定义

云从自然语言中的基本语言值切入,研究定性概念的量化方法,具有直观性和普遍性。定性概念转换成一个个定量值,更形象地说,是转换成论域空间的一个个点。这是个离散的转换过程,具有偶然性。每一个特定的点的选取是个随机事件,可以用其概率分布函数描述。云滴的确定度又具有模糊性,这个值自身也是个随机值,也可以用其概率分布函数描述。在论域空间中,大量的云滴构成的云,可伸缩、无边沿、远观有形、近视无边,与自然现象中的云有着相似之处,所以借用"云(Cloud)"来命名这个概念—数据之间的数学转换。

下面给出云模型的定义:

设 X 是一个普通集合 $X=\{x\}$,称为论域。关于论域 X 中的模糊集合 A,对于任意元素 x 都存在一个有稳定倾向的随机数 $y=\mu_x(x)$,叫做 x 对 A 的隶属度。如果论域中的元素是简单有序的,则 X 可以看做是基础变量;如果论域中的元素不是简单有序的,而根据某个法则 f,可将 X 映射到另一个有序的论域 X' 中,X' 中有一个且只有一个 x' 和 x 对应,则 X' 为基础变量,隶属度在基础变量上的分布成为云。

2.2 云模型的主要特点

(1)所描述的概念的数值具有凝聚性。

(2)云的期望曲线服从正态分布,便于反映大量日常的模糊概念。

(3)对于相同的 x ,其隶属于概念的隶属度具有随机性,会在一定的范围内浮动,这恰好反映了不同的人对同一事物看法的差异。

2.3 云发生器的算法

社会和自然科学的各个分支都已经证明了正态分布的普适性。因此,正态云就成为最基本的云,它在表达自然语言中的基本语言值时最为有用,不妨称为语言原子。

一维云发生器的算法:

当概念对应的数域为一维时,正态云发生器的算法如下:

输入:表示定性概念 A 的3个数字特征值 Ex,En,He;云滴数 N。

输出:N 个云滴的定量值,以及每个云滴代表概念 A 的确定度。

算法:

a.生成以 En 为期望值,He 为标准差的一个正态随机数 En';

b.生成以 Ex 为期望值,En' 的绝对值为标准差的正态随机数 x;

c.令 x 为定性概念 A 的一次具体量化值,称为云滴;

d.计算 $y=e^{\frac{-(x-Ex)^2}{2(En')^2}}$;

e.令 y 为 x 属于定性概念 A 的确定度;

f.$\{x,y\}$ 完整地反映了这一次定性定量转换的全部内容;

g.重复a～f,直到产生 N 个云滴为止。

3 云识别算法的构造

3.1 云识别的识别对象和目的

为了突出重点,云识别的识别对象是那些既不能被

D 中所有识别器单一匹配识别也不属于 SELF'的不确定的抗原 x。目的是为了降低对此类 x 识别的伪肯定率和伪否定率。云识别的重点不在于利用 SELF'构造特异性的识别器,而是利用成熟的识别器集合和 SELF'来对上述识别对象 x 进行决策,重点在于识别的方法和过程。

3.2 几个重要的变量分析

重要的变量有:用于联合识别的识别器的数目 n,判定标准 - 期望的阈值 Ex',熵的阈值 En' 和超熵的阈值 He'。这几个变量随环境的变化而变化,由用户决定。如果系统的安全级别高,则需要识别器的数目多,熵和超熵的阈值都较低。如果安全级别低,则取值反之。

3.3 云滴的生成

首先给出识别云模型的一个定义:设论域 X 是抗原全集 $X = \{x\}$。关于 X 中的识别器集合 $D = \{d\}$,对于任意抗原 x 都存在一个有稳定倾向的随机数 $y = \mu_x(x)$,叫做 x 对 D 的隶属度。X 可以看做是基础变量,隶属度在基础变量上的分布成为云。

设云滴为 (j,k),取此云滴的定量值 j 为抗原 x 和识别器 d 之间的 R - 邻域位匹配位数,在几何模型上是 x 和 d 之间的距离。此云滴代表概念的确定度 k 为 x 和 d 的亲和度[6]。

3.4 云识别算法

针对一个外来的不确定性抗原 x,下面是用 n 个识别器对 x 进行云识别的算法:

输入:抗原 x 和识别器集合 D,并且用户根据安全级别给出变量 n, Ex', En', He' 的值。

输出:x 是 SELF 还是 NONSELF。

算法:

1) 获取待识别的抗原 x。

2) 对于识别器集合 D 中的每个识别器 d,循环

a. 用 R - 邻域位匹配算法算出 d 和 x 的 R - 邻域位匹配位数 j,同时根据 j 的大小将 d 从大到小排序;

b. 计算出 d 和 x 的亲和度 k;

c. 取前 n 个识别器并形成 n 个云滴 (j,k)。

3) 根据这 n 个云滴,参照上文中的一维云发生器的构造算法得到 x 对 D 的期望值 Ex、熵 En 和超熵 He。

4) 比较 Ex, En, He 和 Ex', En', He' 的大小关系:

a. 如果 $Ex \geqslant Ex', En \leqslant En'$ 并且 $He \leqslant Ee'$,那么 x 为 NONSELF;

b. 否则 x 为 SELF。

4 云识别算法的实验系统

为了证明云识别能够降低识别的伪肯定率和伪否定率,文中对此进行了实验。

云识别算法的实验,从总体框架上来看,主要有以下 3 个步骤:初始化数据集,计算两种方法对特定抗原 x 识别的伪否定率,通过比较这两种方法的伪否定率来证明云识别的有效性。相对于传统的识别方法,云识别降低了伪否定率。

4.1 初始化数据集

初始化数据集包括随机生成数据集并训练成熟识别器。

(1) 随机生成的初始数据集包括抗原测试集(由 SELF 测试集和 NONSELF 测试集组成),抗原测试集里已知的 SELF'训练集。这里基于下文的分析用 16 位的数字序列的编码方式来表示抗原和识别器[7]。

(2)生成有效成熟识别器集的算法有很多,比如反向选择算法、克隆选择算法或遗传算法等,本实验选用比较普遍的经过实验证明能产生有效识别器集的反向选择算法(Negative Selection Algorithm, NSA)[8]。

构造识别器之前需要确定 SELF 和 NONSELF 的表示方法,即对抗原进行编码。大量的实验证明,R 邻域位匹配方法适用于数字序列的异常系统调用和病毒的检测。而针对网络通讯数据,R 邻域位匹配方法并不适用。因此选用 10 位的 0 - 1 数字序列的编码方式来表示抗原和识别器。产生识别器集的 NSA 如图 1 所示。

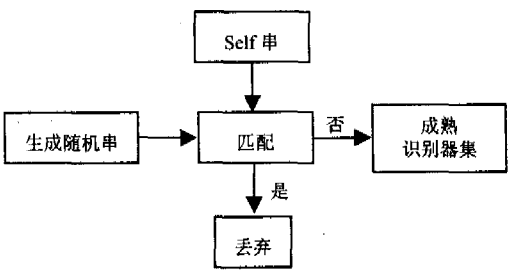

图 1　产生识别器集的 NSA

考虑到实际实验环境及条件限制,参照国外同类实验系统的通行方法,由一个"随机"生成算法产生不同的训练集和测试集,可得到更多的实验数据[9]。然后根据上文介绍的 NSA 训练识别器集。并根据识别器集对抗原的识别率评价识别器集的有效性,丢弃识别率低的识别器集。

4.2 计算伪否定率

由实验目的的分析,待识别的对象是既不属于 SELF'同时也不是能被 D 中的任何一个识别器单独匹配识别的抗原 x,因此需要从初始数据集中将此类抗原 x 从初始数据集里分离出来。

分别计算传统识别方式和云识别对 x 识别的伪否定率:

(1) 传统识别方式的伪否定率:由于这种识别方式将 x 全部当成 SELF(为了简单起见,设系统的安全级别低),因此计算伪否定率只需从数据集中直接计算。

公式为:伪肯定率 = x 中 NONSELF 个数 / x 集合的所有元素个数

(2)云识别的伪否定率:由上文提出的云识别算法构建系统,输入分离出来的抗原 x 和识别器集合 D。得出结果并计算云识别的伪否定率。

计算伪否定率的公式为:伪否定率 = (x 中 NONSELF 个数 - 云识别的 NONSELF 个数) / x 集合的所有

元素个数

4.3 比较结果

针对抗原 x，比较传统识别方法和云识别方法识别它时的伪否定率，从而证明了云识别能够降低此类伪否定率。

5 小 结

将研究不确定性问题的云模型理论引入到人工免疫学的识别过程，借鉴云模型的思想和数学工具，提出了云识别的算法，用以识别不确定性抗原。相对于传统的识别方法对不确定性抗原的识别，云识别降低了伪肯定率和伪否定率。如果有多个识别器能够成功识别很多这样的不确定性抗原，则可将这些不确定性抗原聚类，形成一个多粒度的数据场，从而实现数据场的由微观到宏观的转变。

参考文献：

[1] Chao D L, Forrest S. Information Immune Systems[A]. International Conference on Artificial Immune Systems (ICARIS)[C]. Canterbury, England: University of Kent at Canterbury Press, 2002. 132 - 140.

[2] González F, Dasgupta D, Kozma R. Combining Negative Selection and Classification Techniques for Anomaly Detection [A]. In Proceedings of the Congress on Evolutionary Computation[C]. Honolulu, HI: IEEE Press, 2002. 705 - 710.

[3] 梁意文. 网络信息安全的免疫模型[D]. 武汉：武汉大学，2002.

[4] Forrest S, Hofmeyr S. Immunology as Information Processing [A]. in Design Principles for the Immune Systems and Other Distributed Autonomous System[C]. UK: Oxford University Press, 2001. 361 - 388.

[5] Hofmeyr S, Forrest S. Architecture for an Artificial Immune System[J]. Evolutionary Computation, 1999, 7(1): 1289 - 1296.

[6] 杜 鹢，李德毅. 基于云的概念划分及其在关联采掘上的应用[J]. 软件学报，2001，12(2)：196 - 203.

[7] Forrest S, Perelson A, Allen L, et al. Self - NonSelf Discrimination in a Computer[A]. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy[C]. Los Alamos, CA: IEEE Computer Society Press, 1994.

[8] Dasgupta D, Ji Z, González F. Artificial Immune System (AIS) Research in the Last Five Years[A]. in the proceedings of the international conference on Evolutionary Computation Conference(CEC)[C]. Canberra, Australia: [s. n.], 2003. 8 - 12.

[9] Kim J, Bentley P J. Immune Memory in the Dynamic Clonal Selection Algorithm[A]. in the proceeding of the first International Conference on Artificial Immune Systems (ICARIS)[C]. Canterbury, UK: [s. n.], 2002. 9 - 11.

(上接第 186 页)

就是将统计和结构方法结合的很好范例。

当然，不同方法适用于不同领域的分割要求。信函分拣系统中对于汉字分割的要求很高，需要将几种方法结合分割；而文稿自动输入系统则对于输入速度要求更高，可以要求书写者选用方格纸书写，使汉字间不存在重叠或粘连现象，这种情况下采用垂直投影和模糊匹配的方法就完全可以胜任了。

参考文献：

[1] Chen Y. Research on hand - printed Chinese character recognition[D]. Beijing: Tsinghua University, 1997.

[2] 丁晓青. 汉字识别研究回顾[J]. 电子学报，2002，30(9)：1364 - 1368.

[3] 高彦宇，杨 扬. 无约束手写体汉字切分方法综述[J]. 计算机工程，2004，30(5)：144 - 146.

[4] Lu Y. Machine Printed Character Segmentation - An Overview[J]. Pattern Recognition, 1995, 28(1): 67 - 80.

[5] 朱 锴，赵宇明，吴 越. 一种离线手写体汉字切分的自适应算法[J]. 计算机工程与应用，2004(6)：47 - 50.

[6] 陈 强，娄 震，杨静宇. 非限定手写汉字的分割研究[J]. 南京理工大学学报，2004，28(1)：95 - 98.

[7] Zhao Shuyan, Chi Zheru, Shi Penfei, et al. Two - stage segmentation of unconstrained handwritten Chinese characters [J]. Pattern Recognition, 2003, 37: 145 - 156.

[8] Liang Z, Shi P. A metasynthetic approach for segmenting handwritten Chinese character strings[J]. Pattern Recognition Letters, 2005, 26: 1 - 14.

[9] Casey R G, Lecolinet E. A Survey of Method and Strategies in Character Segmentation[J]. IEEE Trans PAMI, 1996, 18(7): 690 - 706.

[10] 王琳琬，杨 扬，颉 斌，等. 基于连通域单元和穿越算法的汉字切分[J]. 信息技术，2004，28(4)：30 - 35.

[11] Tseng Lin Yu, Chen Rung Ching. Segmenting handwritten Chinese character based on heuristic merging of stroke bounding boxes and dynamic programming[J]. Pattern Recognition Letters, 1998, 19: 963 - 973.

[12] 王 嵘，丁晓青，刘长松. 基于笔画合并的手写体信函地址汉字切分识别[J]. 清华大学学报，2004，44(4)：498 - 502.

[13] Lu Zhongkang, Chi Zheru, Wan - Chi Siu, et al. A background - thinning - based approach for separating and recognizing connected handwritten digit strings[J]. Pattern Recognition, 1998, 32: 921 - 933.

[14] 魏湘辉，马少平. 基于凸包像素比特征的粘连汉字切分[J]. 中文信息学报，2005，19(1)：91 - 97.

[15] Gonzalez R C, Woods R E. Digital Image Processing[M]. Boston: Addison - Wesley, 1992.

[16] Tseng Yi - Hong, Lee His - Jian. Recognition - based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm[J]. Pattern Recognition Letters, 1999, 20: 791 - 806.