

一种检测器的标识学习和优化算法

池 静, 杨振宇, 张 婷

(河北工程大学, 信息与电气工程学院, 河北 邯郸 056038)

摘 要: 为了检测出入侵检测中异常数据的类型以及解决成熟检测器的聚类问题, 提出一种检测器的标识学习和优化算法. 该算法首先对初始化的成熟检测器, 以每个非己抗原为中心进行聚类学习, 得到每个类内检测器的标识, 用以识别异常数据的类型; 然后在经过 N 代之后, 对成熟检测器集中的重叠检测器进行变异和删除操作, 来提高在非己空间的覆盖率. 实验结果表明, 该算法在提高检测器性能的同时, 能够检测出异常数据的类型.

关键词: 入侵检测; 人工免疫; 检测器优化; 聚类; 类型

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2013)08-0019-04

An Algorithm of Identification Learning and Optimization for Detectors

CHI Jing, YANG Zhen-yu, ZHANG Ting

(School of Information & Electronic Engineering, Hebei University of Engineering, Handan 056038, China)

Abstract: In order to detect the types of abnormal database in intrusion detection and to solve the clustering problem of mature detectors, an algorithm of identification learning and optimization for detectors is proposed. Firstly, each initialization mature detector is clustered as the center of non-self antigens to obtain the label of every detector in class, which identifies the types of abnormal database. Then, variation and deleting are operated on the overlapping detectors which are in the set of mature detectors to improve the coverage of non-self space after N -generation. Experimental results show that the proposed algorithm is able to improve the detection performance of the detectors as well as indicating the type of abnormal database.

Key words: intrusion detection; artificial immune; detector optimization; clustering; type

1 引言

入侵检测系统 (Intrusion Detection System, IDS) 是一个监视网络系统中任何非法访问和恶意行为的一个安全系统, 它从各种网络资源中采集信息, 并且分析可能的攻击行为. 从检测机制的角度, 入侵检测可以分为异常检测和误用检测两种类型. 异常检测是在假设入侵行为是异常行为的前提下, 利用已知的系统行为特征和正常用户检测新的未知的攻击行为. 其误报率较高, 这是因为对系统行为特征和用户行为描述是否完备, 此外, 有些和正常行为很相近的入侵行为也会导致较高的误报率. 误用检测是先将攻击的先验知识建立模式, 然后将检测到的行为与已有模式进行匹配, 通过此过程来检测攻

击行为. 相比异常检测, 它对未知的人侵行为或者已知入侵行为变异的检测率较低.

基于人体免疫系统和网络入侵检测系统的相似性, Dasgupta^[1]、Forrest^[2]、Kim^[3] 等人将人体免疫系统应用于网络入侵检测中, 以后有学者进行了深入研究^[4-5]. 人工免疫系统使用有限的抗体 (即检测器) 去检测相对无限的抗原 (即入侵攻击行为), 它将行为分为两类, 分别是有区别的正常行为 (自体) 和异常行为 (非自体). 基于自体比较和基因库的随机生成算法^[6]及改进算法是目前比较常用的检测器生成算法. 这些算法生成的检测器能检测出正常或者异常两种状态的数据, 并且得到较好的检测率, 但是不能识别出异常数据的类型.

本文通过以非己抗原为中心对抗体进行聚类学

习后,能够识别出该非己抗原的类型,即能标识出异常数据的类型.同时提出经过若干代之后,对成熟检测器集进行优化,对浓度较高的类内部分检测器进行变异和删除,保证检测器的多样性.实验表明该方法生成的检测器检测效果较好,能较好地识别异常数据种类.

2 检测器的标识聚类学习过程

在传统的动态克隆选择算法中,经过 T 代耐受成功的未成熟检测器变成了成熟检测器,这些成熟检测器包含了非己的特征^[7],但是由于检测器的分布是随机的,因此非己抗原的分布情况并没有反映出来,通过原算法产生的检测器只能检测出待检测的数据是自己或非己,并不能确定非己数据所属的数据型.应用到入侵检测时,只能检测数据的异常或正常,不能检测出具体的攻击类型.本文通过非己对成熟检测器的刺激作用,实现对检测器的聚类操作,得出每个类的标识,即非己的类型.在检测阶段,通过计算检测器与异常数据的匹配程度,得出该异常数据的类型.

首先按照欧式距离公式计算出各个非己抗原和成熟检测器集合每个检测器之间的亲和力;以每个非己抗原为中心,按照每个检测器同中心点亲和力的大小把检测器空间划分成 A, B, C 三个不同的区域.其中 A 区与检测器最近, B 区次之, C 区最远.假设 A 区半径为 R_1 , C 区半径为 R_2 .对于 A 区内的检测器,他们同非己抗原的欧氏距离最近,和该抗原的亲和力最大,因此 A 区内的检测器不进行变异.对于 C 区内的检测器,同非己抗原的距离最远,这类抗原对检测器的刺激作用最小,因此对其忽略,同样对 C 区内的检测器不做任何改变.而 B 区中的检测器同非己抗原的距离适中,可以通过变异操作提高和非己抗原的亲和力,故对 B 区中的检测器进行复制和删除操作.

复制和删除操作

A 区或者 B 区中没有检测器,这表明检测器集合中没有找到同非己抗原发生匹配的检测器.出现这种情况的原因可能有两种:一种是输入的抗原有可能是自己元素,另一种是检测器集合中没有能够检测该种类型抗原的检测器.为了解决这种情况,首先将输入系统的非己抗原同自体集中的个体进行逐一比较,如果和其中的一个向量相匹配则判定为自己模式,则说明该非己抗原无效,保持检测器集合不

变.若不能同任何自己发生匹配,说明检测器集合中不存在能够检测该非己抗原的检测器,则在区域 A 内直接克隆该非己抗原自身,生成一个检测器,并将其加入到成熟检测集合中去,并在区域 C 中随机选取一个的检测检测器并将之删除,以保证检测器集合的数量不至于过多.

具体步骤如图 1 所示.

(1)得到该时期的非己抗原集合.

(2)利用欧氏距离公式计算每个检测器和该非己抗原的亲和力.

(3)按照每个检测器同非己抗原亲和力的大小将检测器空间分成三个区域,分别为 A, B, C.

(4)仅对在 B 区中的检测器进行有向导变异,变异后得到新的检测器替代先前的检测器.变异因子为 a .

(5)对 A 区和 B 区进行遍历,如果没有检测器落入到这两个区域,则进行复制和删除操作.

(6)重复上面过程,直到所有的非己抗原都输入完毕.

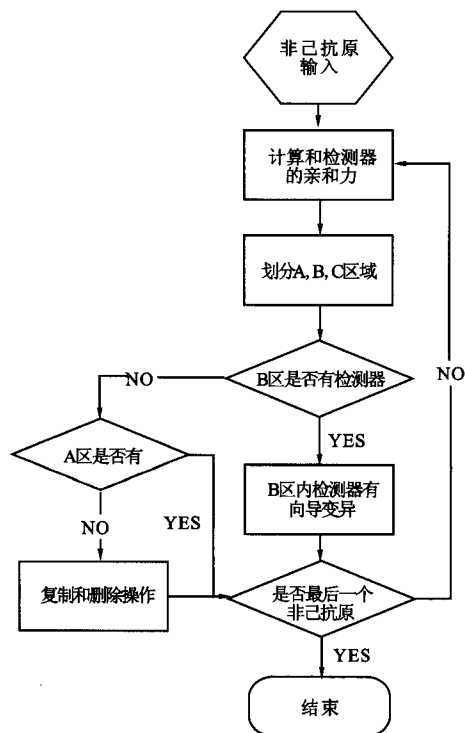


图 1 检测器的标识聚类学习过程流程图

3 成熟检测器优化算法

在传统的基于人工免疫的入侵检测系统中,成熟检测器是随机生成的未成熟检测器经过 T 代耐受成功生成的,对于问题的解决(入侵的检测)没有

预见性,而且使用基因库等的检测器生成算法,由于基因库本身的聚类问题,都会造成成熟检测器集合存在过度聚类情况,即存在重叠检测器,影响检测器的多样性,这些冗余的检测器不仅占用宝贵的检测器空间,而且影响到入侵检测系统的整体性能^[8]. 同时在以后遇到协同刺激信号否定时需要删除相似的检测器多造成资源很大浪费,因此有必要对成熟检测器集进行优化.

代数太少重叠情况不明显没有必要进行此操作,因此在经历过 N 代之后,再对成熟检测器集进行聚类. 当此类检测器浓度达到一定程度时,对该类部分检测器进行变异并删除,加入到未成熟检测器中进行自体耐受. 达到消除部分冗余,保持抗体的多样性,进而提高检测器对非己空间的覆盖率的目的.

聚类是指按照事物的某些属性,在没有事先给定聚类的数目和结构的情况下,把事物聚集成类,使得类内相似性尽量大,类间相似性尽量小. 聚类是数据挖掘的主要方法之一,同时也是统计数据分析的一个常见技术. 聚类算法的主要分类有:层次的方法;基于密度的方法;基于网格的方法;基于模型的方法等等.

层次聚类有两种分类情况. 第一种是凝聚方法,即自底向上思想. 一开始将每个对象作为单独的一组,然后根据同类相近,异类相异的原则,合并对象,直到所有的组合成一个,或达到一个终止条件为止. 另一种是分类方法,自顶向下思想. 一开始将所有的对象置于一类,在迭代的每一步中,一个类不断地分为更小的类,直到每个对象在单独的一个类中,或达到一个终止条件. 层次聚类的优点是事先不用先定好类的个数. 计算类间距有三种方法,分别如下:

(1)最小距离法,计算两类间的最小距离,使用下式计算:

$$\text{dist}(P, Q) = \min \|x_P, x_Q\| \quad (1)$$

式中, $i = [1, n_P]$, $j = [1, n_Q]$.

(2)类平均距离法,类中所有样本点的平均距离,使用下式计算:

$$\text{dist}(P, Q) = \|\bar{x}_P - \bar{x}_Q\| \quad (2)$$

其中, $\bar{x}_P = \frac{1}{n_P} \sum_{i=1}^{n_P} x_{P_i}$, $\bar{x}_Q = \frac{1}{n_Q} \sum_{i=1}^{n_Q} x_{Q_i}$.

(3)差离平方和法,计算公式如下:

$$\Delta_{\text{dist}}(P, Q) = \text{ESS}(PQ) - \text{ESS}(P) - \text{ESS}(Q) \quad (3)$$

其中,

$$\text{ESS}(P) = \sum_{i=1}^{n_P} |x_{P_i} - \bar{x}_P|^2, \quad \text{ESS}(Q) = \sum_{i=1}^{n_Q} |x_{Q_i} - \bar{x}_Q|^2.$$

在本文中,使用类平均距离计算类间距,使用不相关系数作为终止条件. 因此当该类的不相关系数超过阈值 T , 停止聚类. 聚类完成之后检测每个类的浓度,如果超过阈值 Y , 则对随机 $n\%$ 的个体变异. 从而改进改进检测器对非自体空间的覆盖率.

具体步骤如下:

Step1: 检测是否满足算法执行条件,即与上次算法执行抗体代数间隔与阈值 T_1 比较,若大于阈值 T_1 执行算法.

Step2: 对检测器集合进行层次聚类,使用不相关系数为终止条件. 若满足不相关系数小于 T_2 即执行合并操作.

Step3: 对聚类后的检测器集合,检查类的浓度(类内对象个数/对象总数)是否大于阈值 T_3 , 大于即执行 Step4 操作.

Step4: 随机对 $n\%$ 的个体进行变异,然后删除之,放入未成熟检测器.

4 仿真结果及分析

阈值 R_2 影响检测器聚类问题,如果 R_2 过大,则会导致需要变异的检测器过多,类型之间会出现明显的边界,即出现干扰情况;而如果 R_2 过小,则会无法区分类型之间边界,聚类效果会很差. 变异因子 α 影响聚类学习效果. 如果 α 过小,会使检测器的变异程度小,而取值过大,又会使得相邻的非己抗原之间的干扰过大. 参考相关文献和经过多次试验,设置 $R_2 = 0.09$, $R_1 = 0.03$, $\alpha = 0.15$. 其他参数为 $T_1 = 120$, $T_2 = 1.2$, $T_3 = 0.3$.

实验环境: CPU Core i5 2.8 GHz, 内存 2 GB, 操作系统为 Windows 7, 编程工具 Matlab R2010a, 数据集采用 KDD CUP99. 从 KDD CUP99 的训练样本中选取 4 000 条数据,其中 3 000 条正常数据,1 000 条包括 DOS、R2L、U2R 及 PROBE 等入侵数据. 测试集从余下样本集中选取 1 200 条正常数据,800 条入侵数据,其中 DOS 300 条, R2L 200 条, U2R 50 条, PROBE 250 条.

将使用了成熟检测器优化的算法与原克隆选择算法检测率比较,比较结果如表 1 所示.

表 1 不同代数下两种算法检测率 %

代数	原克隆选择算法	本文算法
200	32	35
400	51	53
600	65	68
800	72	77
1 000	76	83

从表 1 可以看出,随着代数的增加,本文提出算法的检测率也随之提高.另外,和原克隆选择算法相比,本文提出算法的检测率有所提高.

对异常抗原类型的检测执行如下步骤:找出测试数据向量和检测器集中亲和力小于阈值 Y 所有检测器,排序后选取最小的 K 个,通过检查各检测器的类型得出该数据的类型.其中 $Y=0.06$ $K=3$.

代数为 5 000 时,异常数据类型检测效果如表 2 所示.

表 2 各类型异常数据检测率

	NORMAL	DOS	R2L	U2R	PROBE
检测率/%	96.30	98.60	96.55	86.38	97.92

从表 2 看出:DOS 类型异常数据检测率最高,U2R 类型检测率最低.各个类型的检测率都达到了较好的效果,因此本文提出的算法可以很好地识别异常数据类型.

5 结束语

本文提出在基于人工免疫的入侵检测系统中,通过对成熟检测器进行标志聚类学习,使检测器具具备非己抗原的类型特征,同时对成熟检测器集中的过度聚类问题进行了解决,优化了检测器集,提高了检测器对非己空间的覆盖率.实验表明达到了较好

的检测率同时能够对异常数据的类型有很好识别.

参考文献:

- [1] Dasgupta D, Gonzalez F. An immunity-based technique to characterize intrusions in computer networks [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(3): 281—291.
- [2] Stephanie F, Steven A, Anil S. Computer innunology [J]. Communications of the ACM, 1997, 40(10):88—96.
- [3] Kim J, Bentley P J. Towards an artificial immune systemfor network intrusion detection;an investigation of dynamic clonal selection[C]//Evolution Computation CEC 2002. Hawaii,2002:1015—1020.
- [4] 符海东,王枫.基于数理统计及人工免疫的入侵检测模型[J].计算机工程与设计,2008,29(12):3037—3039.
- [5] 邓荣.基于人工免疫的网络入侵检测模型[J].煤炭技术,2011,30(4):105—107.
- [6] Forrest S, Perelson A, Cherukuri R. Self-nonsel self discrimination in acomputer[C]// Proc. of 1994 IEEE Computer Society Symposium on Research in Security and Privacy. Los Almitos, CA, USA: IEEE Computer Society, 1994: 202—212.
- [7] 姜懿庭.基于人工免疫系统的检测器生成算法改进[J].云南民族大学学报:自然科学版,2012,21(2):141—144.
- [8] 肖军弼,黄波涛.基于改进基因库的检测器生成算法[J].计算机系统应用,2011,20(4):251—253.

作者简介:

- 池 静 女,(1973—),硕士,副教授.研究方向为数据挖掘.
杨振宇 男,(1988—),硕士研究生.研究方向为网络安全.
张 婷 女,(1986—),硕士研究生.研究方向为网络安全.

(上接第 18 页)

参考文献:

- [1] 刘鹏.云计算[M].北京:电子工业出版社,2010
- [2] 王鹏.云计算的关键技术与应用实例[M].北京:人民邮电出版社,2010
- [3] Tom White,周敏,曾大聘,等. Hadoop 权威指南[M].北京:清华大学出版社,2011
- [4] Zhang Hong, Lu Song feng, Zhao You qiao, et al. Research on file—slice model for data security storage [J]. Computer Engineering and Application, 2012, 48 (18): 66—70.

- [5] 王鹏.走进云计算[M].北京:人民邮电出版社,2009: 221—224.
- [6] Huang Xiaotao, Li Zhiyong. Replication strategy based on file—partition for P2P network [J]. Application Research of Computers,2008, 25(1): 171—173.

作者简介:

- 钱进进 女,(1988—),硕士研究生.研究方向为网络与信息安全.
凌 捷 男,(1964—),博士.研究方向为网络与信息安全.