基于改进基因库的检测器生成算法①

肖军弼，黄波涛

(中国石油大学(华东) 计算机与通信工程学院，东营 257061)

摘 要：在基因库生成检测器算法中，一般是把被删除的记忆检测器进行基因突变后的基因或非自体集样本加入到基因库中来初始化并更新基因库。经过若干代之后，在基因库中会出现一些相似性比较大的基因，形成基因的聚类现象。通过定期的对基因库进行聚类，变异，约减，提高成熟检测器对入侵的检测多样性。实验结果表明，该方法是有效的，能在快速生成检测器的同时，提高对未知入侵的检测能力。

关键词：入侵检测；人工免疫；检测器生成；基因库

Detectors Generating Algorithm Based on Modified Gene Library

XIAO Jun-Bi, HUANG Bo-Tao

(School of Computer and Communication Engineering, China Petroleum University (East China), Dongying 257061, China)

Abstract: The gene always comes from mutating the deleted memory detectors or no-self samples in the algorithm of generating detectors based on gene library. But after some generations, there will be some similar gene in the gene library. This is the phenomenon of cluster of gene. By using clustering, mutating, subtracting to gene library, it can increase the diversity of mutated detectors. The experimental results show that the algorithm can generate detectors more quickly and improve the ability of detection system to discover unknown intrusion.

Keywords: intrusion detection; artificial immune; detectors generating; gene library

1 引言

受到人体免疫系统的启发，Dasgupta、Forrest、Kim 等人将人工免疫系统引入到网络入侵检测当中来，并做了深入的研究。人工免疫系统将行为分为有差异的正常行为和异常行为，并用有限的抗体(检测器)来检测相对无限的抗原(入侵行为)。因此，生成检测器的效率和检测器对入侵的识别成为了人工免疫系统应用在入侵检测中的核心问题。比较常用的检测器生成算法是随机生成算法[1]及其改进算法。这类算法通过随机生成字符串来产生未成熟检测器，检测器会均匀的覆盖自体空间和非自体空间。但是在网络数据中，正常数据流量要比入侵数据流量高，因此会产生很多不合格的检测器，浪费大量的时间和运算，效率比较低。人体免疫系统使用基因库来产生抗体，通过克隆，高频变异激活的抗体细胞，产生更多的变种基因，使抗体向未知的抗原空间搜索[2]。人体免疫系统的这种特性启发人们使用基因库来产生检测器。基因库一般是通过对被删除的记忆检测器进行高频变异[2]或者利用非自体集合样本[3]形成的。随着基因库内基因的增加，在基因库中出现相似的基因形成聚类。簇内基因过多，在随机选择基因时，大簇选择的概率就比较高，产生的检测器的多样性就受到影响。这样有限的检测器不能尽量广的覆盖非自体空间，从而导致检测率不高，同时也会增加算法的空间开销。文献[4]提出了一种分布式的基因库生成检测器算法，如果基因添加到基因库的时间超过设定的阈值则删除该基因，这样只考虑存活时间就会把一些优秀的基因也删除掉。本文提出经过若干代之后，对基因库进行聚类，对浓度较高的簇内基因进行高频变异和删除，从而保证基因库的多样性。通过实验表明该方法生成的检测器检测效果较好。

① 收稿时间:2010-09-09;收到修改稿时间:2010-10-15

2　基于否定选择的随机检测器生成算法

Forrest 等人在 1994 年提出了否定选择的思想，并将随机检测器生成算法引入其中[1]。该算法首先定义自体集合，再随机生成候选检测器，将该检测器与自体集中元素比较，如果匹配成功，则删除该检测器，否则成为合格检测器。

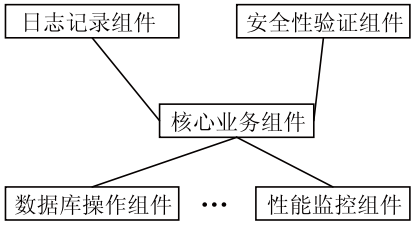

图 1　随机生成检测器算法模型

3　基于基因库的检测器生成算法

随机生成算法产生的检测器能均匀的覆盖自体集与非自体集。但是在网络数据中，一般正常数据流量往往要比非正常流量(入侵数据)高出很多，因此会随机生成很多不合格的检测器。在人体免疫系统中，免疫细胞是由基因库中随机选出的不同基因成分串联而成。Perelson 等人根据人体的这种特性构建出了骨髓模型[5]。在骨髓模型中，基因库中的基因来源于人体已经产生的免疫细胞，从而利用已有的免疫结果来指导新免疫细胞的生成。将这种原理应用在入侵检测中来，就可以利用已经产生的检测器反过来指导新检测器的生产，能提高检测器生成的效率。

3.1 使用基因库生成未成熟检测器

图 2 是一个典型的使用基因库生成检测器的人工免疫模型。基因库中的基因来源于被删除的记忆检测器或者非自体集合样本，从每个基因库中选取一个基因最后串联成为一个新的检测器。

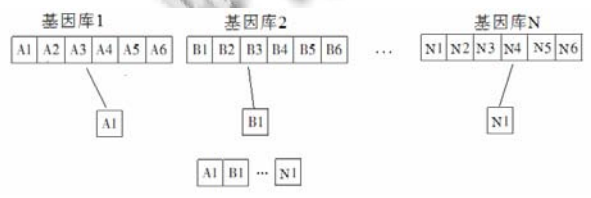

图 2　基因库生成检测器模型

3.2 基因库的处理

目前构建基因库的方法有对被删除的记忆检测器进行突变[2]，或使用非自体集合中的样本[3]。经过一段时间之后，随着被删除记忆检测器的增多，产生的基因会出现聚类现象。文献[3]采用非自体集合样本对基因库进行初始化，将所有非我样本输入基因库，基因库的每一行就是一个非我样本。由于非我样本比较多，并且出现同一类攻击会有很多个非我样本，这样造成基因库的冗余比较多，也会产生聚类现象。在生成检测器的时候是随机从基因库中选出的，当基因库中出现聚类时，大簇内的基因就会被以更高的概率选出。这样产生检测器的多样性就会受到影响，使检测器对非自体集合的覆盖范围缩小，进而影响检测效率，同时冗余的基因也会增加算法的空间开销。本文提出在经过若干代之后，对基因库进行聚类，当一个簇浓度大于预定阈值 c 的时候，首先计算聚类的半径，对于小于 0.5 倍半径内的基因，随机选择 n%个个体删除；对于 0.5 倍半径到半径之间的基因，随机选择 n%个个体进行基因高频突变，使其偏离原来的聚类，向其他非自体空间进行搜索，发现新的抗原。模型如图 3 所示。

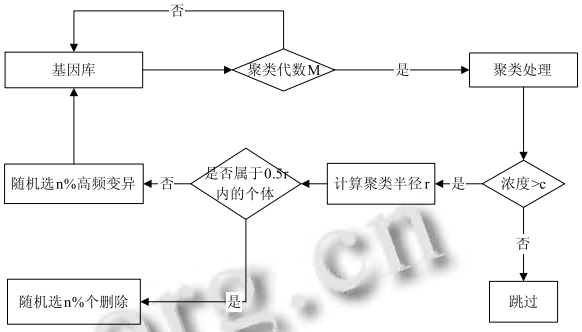

图 3　基因库处理过程

算法描述：

簇的浓度：该簇内的基因的个数/基因总数

If(时间间隔>T)

　　对基因库进行聚类，得到簇的个数 n;

　　int i=0;

　　While(i<n){

　　　　If(该簇浓度>c){

　　　　　　计算簇的半径 r;

　　　　　　对于 0.5r 内的基因， 随机选择 n%个删除；

　　　　　　对于 0.5r 到 r 之间的基因，随机选取 n%个基因进行高频变异；

　　　　　　将变异后的基因添加到基因库中；

```
            }
        }
```

4 实验与分析

实验环境：CPU P4 2.8GHz，内存 1024MB，操作系统为 WindowsXP，编程工具 Visual C++，数据集采用入侵检测领域比较权威的 KDDCUP99。从 KDDCUP9910%的训练样本中选取 8000 条数据，其中 5000 条正常数据，3000 条包括 DoS、R2L、U2R 及 PROBE 等类型的入侵数据。测试集从剩下 10%的样本集中选取 5000 条正常数据，3000 条入侵数据，其中，1500 条 DoS、R2L、U2R 及 PROBE，1500 条 teardrop，portsweep，httptunnel 等未知入侵数据。未成熟检测器的耐受期 T=30，成熟检测器的激活阈值 A=10，成熟检测器的生命周期 L=40，循环代数 N=50，基因库的变异率 mu=0.1，基因库聚类代数 M=100。将约减基因库的算法添加到基于基因库产生检测器的动态克隆算法中去，与直接利用基因库产生检测器的克隆选择算法进行比较。结果如图 4 所示：

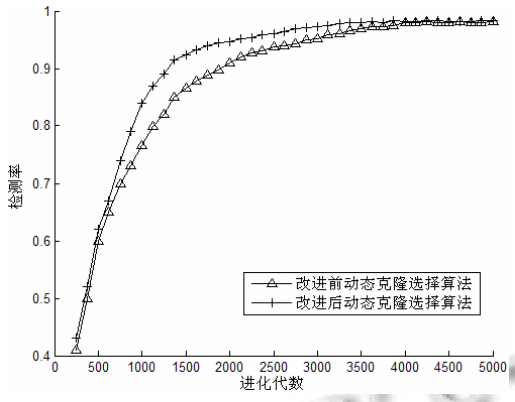

图 4 测试结果对比图

从图 4 中可以看出，在 700 代之前改进后的算法和改进前的算法的检测率没有太大的差别，这是由于这一阶段产生的被删除记忆检测器较少，由他们生成的基因库里的基因还没有形成太多的聚类现象。而在 700 代 3500 代之间检测率比改进前有明显提高，这是由于被删除的记忆检测器在基因库中形成聚类，经过变异处理后，使得检测器的覆盖范围增大，提高了检测效果。而在 3500 之后，检测效果较改进前的算法没有明显的提高，因为随着被删除的检测器增多，基因库中的基因在空间中的分布情况比较均匀，变异算子和删除算子对提高基因在空间搜索能力上的作用减弱。

5 结论

在人工免疫系统中，利用基因库生成检测器的算法比随机算法能更高效的产生检测器，但是随着基因库内基因的增多，这些基因会出现聚类的现象。在随机选择基因的时候，浓度较高的簇内个体被选择的概率要大，这样生成的检测器不能有效的覆盖非我空间。对聚类的基因进行高频变异，并删除相似度较高的基因，使得基因扩大搜索范围，同时降低聚类群体的浓度。利用经过改进基因库生成的检测器能有效的生成检测器，同时提高检测效率。

参考文献

1 Forrest S, Perelson A, Cherukuri R. Self-nonself discrimination in acomputer. Proc. of 1994 IEEE Computer Society Symposium on Research in Security and Privacy. Los Almitos, CA, USA: IEEE Computer Society, 1994: 202−212.

2 Kim J, Bentley PJ. Immune memory and gene library evolution in the dynamic clonal selection algorithm.Journal of Genetic Programming and Evolvable Machines, 2004,5 (4):361−391.

3 王保义,王玮,王蓝婧.人工免疫中一种新的基因库初始化方法.计算机工程与应用,2007,43(21):126−128.

4 葛丽娜,钟诚.基于人工免疫入侵检测检测器生成算法.计算机工程与应用,2005,23:149−152.

5 Perelson A S, Hightower R, Forrest S. Evolution and Somatic Learning in V-Region Genes. Research in Immunology, 1996,147(4):202−208.