利用聚类改进动态克隆选择算法的自体纯净性问题①

肖军弼 季翠翠 (中国石油大学(华东) 计算机与通信工程学院 山东 东营 257061)

摘 要: 动态克隆选择算法应用于入侵检测的过程中，经过记忆检测器和成熟检测器检测后的剩余抗原被直接作为自体供未成熟检测器耐受，但这些剩余抗原并非完全是自体，有可能隐含新型攻击。为此提出利用聚类分析技术进行改进，先用聚类算法将剩余抗原分成大、小簇，然后分析小簇中的数据，发现其中隐含的新型攻击，并及时更新记忆检测器集和自体集。实验结果表明，加入聚类分析的动态克隆选择算法能够增强检测系统发现未知入侵的能力。

关键词: 动态克隆选择算法；入侵检测；剩余抗原；聚类

Using Clustering to Improve Self-Purity of Dynamic Clonal Selection Algorithm

XIAO Jun-Bi, JI Cui-Cui

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

Abstract: In the intrusion detection process of dynamic clonal selection algorithm, the antigens detected by memory detectors and maturity detectors are directly considered as self immature detectors to be tolerated. But there may be new attacks hidden in these antigens. To solve this problem, a new idea with clustering analysis is proposed. The clustering algorithm cluster remaining antigens then analyzes data existing in small cluster, finds hidden attacks and update memory detector set in time. The experimental results show that the dynamic clonal selection algorithm with clustering analysis can enhance the detection system's ability to discover unknown intrusions.

Keywords: dynamic clonal selection algorithm; instrusion detection; remaining antigen; cluster analysis

1 引言

入侵检测系统是网络完全防御体系的一个重要组成部分，通过监视网络或系统资源，寻找违反安全策略的行为或攻击迹象，为网络系统提供保护，并且能对某些突发状况进行预防。生物免疫系统与入侵检测系统具有本质的相似性，基于人工免疫的入侵检测系统模拟了生物免疫系统的耐受性、分布式、自组织、自学习、自适应等优良特性来保护计算机免受非法入侵[1]。Kim、Bentley 在文献[2]中首次提出动态克隆选择算法，但将其应用于实际入侵检测中主要存在以下问题：首先其实现需要大量的协同刺激；另外，要求用于否定选择的自体集必须是纯净的，不存在噪音数据。文献[3]通过删除能检测到自体的记忆检测器，并利用记忆检测器中存在非我信息，对被删除的记忆检测器进行基因突变，产生未成熟检测器，在一定程度上解决了协同刺激过多的问题。文献[4]采用小生境技术对成熟检测器进行进化，鼓励不同种群的检测器进行交叉和变异，增加了成熟检测器的多样性，提高了对未知入侵的检测能力。但在这些改进的动态克隆选择算法的实现过程中，都没有很好的解决自体集的纯净性问题。经过记忆检测器和成熟检测器检测后的抗原(本文称为剩余抗原)被作为自体供未成熟检测器耐受，但这些剩余抗原中有可能隐含某种新型攻击，被作为自体耐受，必然影响检测率。

在前人研究的基础上，对动态克隆选择算法作进一步改进，提出利用聚类算法，对经过记忆检测器和成熟检测器检测之后的剩余抗原进行聚类，然后分析聚类形成的小簇数据，进而迅速发现新型攻击和新自

① 收稿时间:2009-09-07;收到修改稿时间:2009-11-01

体，及时更新记忆检测器集和自体集，能够提高自体的纯净度，进而提高有效检测器的覆盖范围，增强系统检测未知入侵的能力。

2　基于聚类分析的动态克隆选择改进算法

2.1 动态克隆选择算法的自体纯净性问题

文献[5]建立了基于扩展的动态克隆选择算法的入侵检测模型，该模型包括三个重要模块：记忆检测器检测模块，成熟检测器检测模块和未成熟检测器检测模块，分别对应免疫系统的记忆细胞集合，成熟细胞集合和未成熟细胞集合。整个算法的实现分为三个阶段：第一阶段是未成熟检测器经自体耐受进化为成熟检测器阶段；第二阶段是成熟检测器进化阶段，为自学习阶段，在检测过程中通过克隆选择进化为记忆检测器；第三阶段是从记忆检测器产生到算法终止，免疫系统各部件产生完毕，进行实际环境中的检测，其检测流程见图 1。

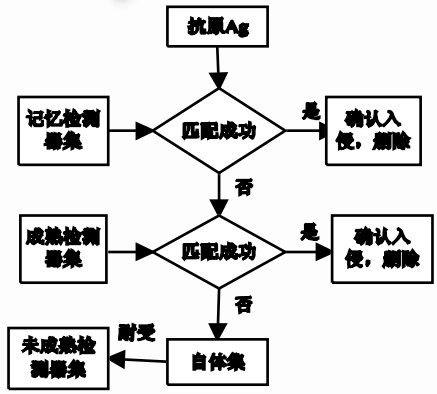

图 1　检测器对抗原的检测过程

抗原首先和记忆检测器进行匹配，如果匹配成功，经外部协同刺激确认是入侵数据后，删除该抗原；剩下的抗原再和成熟检测器匹配，如果匹配成功，同样经外部协同刺激确认是入侵数据后删除；将最后剩下的抗原即剩余抗原作为自体供未成熟检测器耐受。由于有效检测器的覆盖范围是有限的，只是整个非我空间的一个子集，并不能完全覆盖整个非我空间，不能被记忆检测器和成熟检测器匹配的抗原并非都是自体，如果其中隐含某种新型攻击而被作为自体供未成熟检测器耐受，这样系统就不会产生能检测到该入侵的有效检测器，从而影响系统检测未知入侵的能力，造成漏报。

2.2 聚类分析简介

聚类是一个将数据集划分为若干组或类的过程，它将相似的数据划分到同一个聚类中，而将不相似的数据划分到不同的聚类。使得同一个组内的数据对象具有较高的相似度，而不同组中的数据对象则不相似。聚类分析是一种常用的非监督异常检测方法，其实现基于两个前提：一个是在网络数据中正常数据的数量是远远大于入侵数据的；另一个前提是入侵数据和正常数据之间存在很大的区别[6]。现有的主要聚类算法有划分方法、层次的方法、基于密度的方法、基于网格的方法和基于模型的方法等。在聚类能够较好地划分正常和异常数据的前提下，可以将聚类的结果根据大小簇的定义划分成大簇和小簇。正常数据所在的簇应该是大簇，而数量较小的异常数据所在的簇应该是小簇。对聚类后生成的所有簇，计算其成员个数，当某个簇中成员个数小于给定的阈值时，就认为该簇是小簇，其中的数据是异常数据。

2.3 改进的动态克隆选择算法

为解决动态克隆选择算法的自体纯净性问题，引入聚类分析技术，对通过记忆检测器和成熟检测器检测后的剩余抗原进行聚类，由于正常数据的数量远远大于入侵数据，而且大部分入侵数据已经被记忆检测器和成熟检测器检测并处理掉，剩余抗原中隐含的入侵数据只占极少数，因此只对包含极少数抗原的小簇数据进行分析，若发现是新型攻击，则将其从剩余抗原中删除并直接添加到记忆检测器集，再将剩下的抗原添加到自体集。

算法 1.　剩余抗原聚类算法

设 R 为聚类半径阈值，P_i 表示第 i 个抗原，$dist(C_j,P_i)$ 为抗原 P_i 到聚类 C_j 的距离，n 为剩余抗原总数，则剩余抗原聚类算法描述如下：

Begin:

初始化聚类集 S 为空集；

随机选取抗原 P_0，建立以 P_0 为中心的聚类加入到 S；

While($\Delta r > \delta$)Do{ //Δr 为前后两次聚类中心的偏移量，δ 为一较小正数

For(int =0; <n; ++){

If($\exists C_j \in S$ and $dist(C_j,P_i) <= R$){ P_i 加入到 C_j 中；}

else{建立以 P_i 为中心的新聚类并加入到 S 中；}

}

重新确定聚类中心;

}

End

算法 2. 大小簇的划分及小簇分析算法

剩余抗原经算法 1 聚类后分成若干簇，对这些簇的成员进行统计，将其分为大、小簇，并将小簇数据交由外部协同刺激处理。具体算法描述如下:

Begin:

For(int j =0; $j < m$; j ++){

Num(C_j);//统计各簇成员个数

If(Num(C_j)< n_0){//n_0为区分大小簇阈值

外部系统仲裁;

If(入侵数据){

从剩余抗原中删除;

入侵模式添加到记忆检测器集合;

} Else{添加到自体集;}

}Else{

添加到自体集; }

}

End

改进的动态克隆选择算法的侵检测模型如图 2。

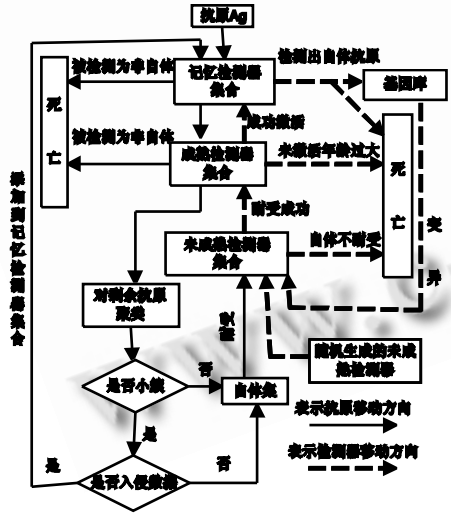

图 2　改进的动态克隆选择算法入侵检测模型

3　实验与分析

实验环境：CPU P4 2.8GHz，内存 512MB，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 KDDCUP99 数据集中挑选 10000 条正常数据(记为 S)和 5000 条包括 DoS、R2L、U2R 及 PROBE 等类型的入侵数据(记为 N)，另外再挑包括 teardrop，portsweep,httptunnel 等 1000 条 N 中未出现的入侵类型(记为 R)。每代输入 200 个抗原，包括从 S 中随机选择的 150 个正常数据和从 N 中随机选择的 50 个入侵数据。系统运行 1000 代以后，向每代抗原中添加 5 条从 R 中随机选择的入侵数据。先后用加入聚类分析的动态克隆选择算法和原动态克隆选择算法进行测试，结果如图 3。

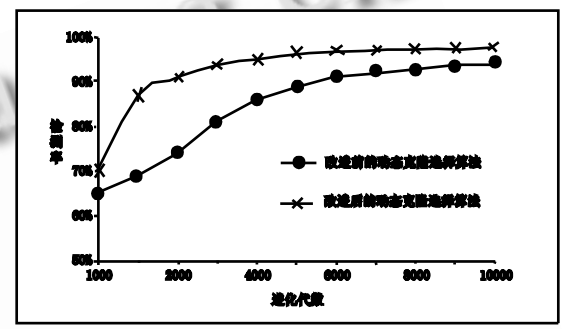

图 3　测试结果比较

从图 3 中可以看出， 在 1000~4000 代之间，加入聚类分析的动态克隆选择算法的检测率明显高于改进前的算法，这是因为改进的算法通过对剩余抗原聚类分析，发现隐含的新型攻击类型，并将其直接添加到记忆检测器集，缩短了二次免疫应答的反应时间。随着进化代数的增加，两算法的检测率之差相对缩小，这是由于动态克隆选择算法的自学习能力，有效检测器的覆盖范围有所提高造成的。但从总体来看，和原动态克隆选择算法相比，聚类分析的加入，提高了入侵检测的检测率。

4　结语

在基于免疫的入侵检测中，有效检测器的生成必须经历一个自体耐受的过程，自体纯净度会影响检测器覆盖非我空间的范围，对动态克隆选择算法中剩余抗原聚类分析，提高了自体的纯净度，有利于增强系统发现未知入侵的能力。虽然利用聚类改进的动态克隆选择算法提高了入侵检测的检测率，但其实现仍然需要大量协同刺激。如何在提高检测率的同时，减少协同刺激的数量是下一步工作的重点。

(下转第 215 页)

(上接第 173 页)

参考文献

1 李涛.计算机免疫学.北京:电子工业出版社, 2004.

2 Kim J, Bentley PJ.Towards an artificial immune system for network intrusion detection:an investigation of dynamic clonal selection. Proc. of the Congress on Evolutionary Computation 2002, Honolulu, 2002. 1015－1020.

3 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.

4 冯艳华,钟诚.动态克隆选择的成熟检测器进化算法.微电子学与计算机, 2007,24(10):181－183.

5 Leonid Poanoy, Eleazar Eskin, Salvatore J Stolfo. Intrusion detection with Unlabeled Data Using Clustering. Proc. of ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001). Philadelphia, PA: November 2001. 5－8.