第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)13—0166—03 文献标识码：A 中图分类号：TP393

基于粗粒度遗传算法的网络入侵检测系统

李 甦，罗安坤
(云南大学信息学院，昆明 650091)

摘 要：分析遗传算法在入侵检测系统中的可应用情况，提出一种基于粗粒度模型遗传算法的网络入侵检测系统。通过对协议特征的分析，找出有可能被非法利用和更改的特征属性，经过组合和编码后构成系统的初始种群，在各个处理器(终端点)并行地进行遗传算法的操作，使种群的进化在所有检测点同时进行，通过迁移相互交流，合理地设计适应度函数，使遗传“基因”的取舍和利用更加合理。实验数据表明，系统的检测率达到90%以上。

关键词：网络入侵检测系统；遗传算法；种群；适应度函数；粗粒度模型

Network Intrusion Detection System Based on Coarse-grained Model Genetic Algorithm

LI Su, LUO An-kun
(College of Information, Yunnan University, Kunming 650091)

【Abstract】This paper puts forward a Network Intrusion Detection System(NIDS) based on coarse-grained model genetic algorithm, after analysing the application of genetic algorithm in intrusion detection system. By analysing the protocol's property, finds out some characteristics which are often lawlessly changed and used, forms the original chromosome by assembling and coding, makes all processor(terminal points)carry genetic process by attributed manner, makes the evolution of chromosomes prosecute in all detection points simultaneity, and all processor can intercommunacate by transplanting too. At the same time, redesign a reasonable fitness function, to let the use of "gene" be more reasonable. Experimental data shows the system's detection ability is above 90%.

【Key words】Network Intrusion Detection System(NIDS); Genetic Algorithm(GA); chromosomes; fitness function; coarse-grained model

1 遗传算法在网络入侵检测系统中的应用

1.1 网络入侵检测系统

入侵检测系统按照所检测的数据来源分为基于主机的入侵检测系统和基于网络的入侵检测系统(Network Based Intrusion Detection System, NIDS)[1]。基于网络的入侵检测系统主要通过对网络上数据包的分析来检测通过网络发起的入侵活动。在入侵检测系统中需要定义入侵特征和用户正常行为轮廓的参数集，为了实现参数集合和属性的最优化，引入了遗传算法来搜索整个度量空间，以适应度评价的方式逐步优化初始度量参数子集，从而得到针对特定检测环境的最优度量集合。

1.2 粗粒度模型遗传算法

遗传算法(Genetic Algorithm, GA)中适应度的计算最费时间，加上不断地产生新的一代，而每一代中又有若干个个体，所以，如何提高遗传算法的运行速度显得尤其重要。由于遗传算法具有内在的并行机制，再加上网络系统规模的扩大，因此将并行遗传算法[2]运用到网络入侵检测系统中，从而构建更优秀的特征库，以期获得更好的检测效果。并行遗传算法的实现方式有3种[2]，其中粗粒度模型(独立型)遗传算法的特点是将整个种群分成多个子种群，各子种群在不同的处理器上相对独立地并发执行操作，不仅独立进行计算适应度，而且独立进行选择、重组交叉和变异操作，并以一定间隔在子群体间交换个体，即各子种群通过交换若干个体以引入其他种群的优秀个体，从而加快满足终止条件的要求。其算法流程如图1所示。

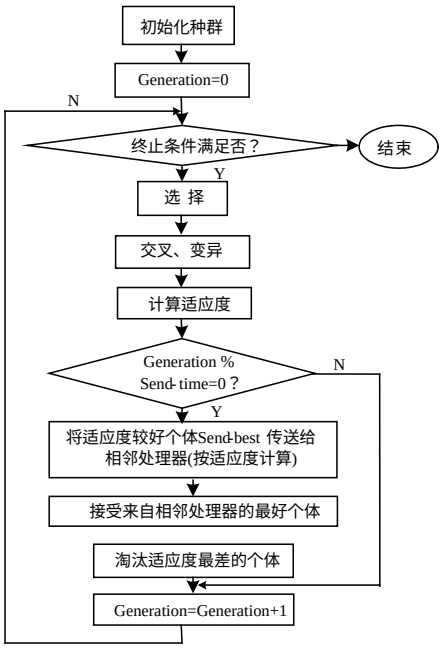

图1 粗粒度模型遗传算法流程

基金项目：云南省教育厅应用基础研究基金资助重点项目(03Z180A)
作者简介：李 甦(1963－)，男，教授，主研方向：计算机安全，信息处理；罗安坤，硕士研究生
收稿日期：2007-08-20 **E-mail：**luoankun@163.com

2　迁移算法和适应度函数的设计

2.1　迁移算子

粗粒度模型遗传算法除了基本的选择、交叉、变异等操作外，还引入了迁移算子，即在进化过程中子群体间交换个体的过程。一般的迁移方法是将子群体中较好的几个个体迁移给其他的子群体。这样就只需要较少的个体评价计算工作量，避免了局部早熟现象的产生，同时也不影响算法的运行速度。

迁移策略采用阶梯式迁移(一传一)策略：当产生新一代的个体后按照算法的要求只将适应度最好的几个个体传送给与自己相邻的一个处理器，并且接受来自这个处理器的最好的几个个体，然后将这些个体同自己的个体同时考虑，淘汰适应度差的个体。这样加强了处理器之间优秀个体的交流，扩大了全局寻优，避免个体陷入局部最优化，扩大了遗传算法的搜索范围。

2.2　适应度函数设计

遗传算法在搜索进化过程中一般不需要其他外部信息，仅用适应度(Fitness)函数[3]来评估个体或解的优劣性，并作为以后遗传操作的依据。

研究表明，个体的适应度值可以取决于共有多少攻击被正确检测出来和共有多少正常使用连接被误检测为攻击，即可以根据检出率和误检率两个参数综合考虑后来设计适应度函数。

设：Xi 为某个个体。

Tc 为总的正常连接(total connect)数目。

Ta 为总的攻击(total attack)数目。

Ra 为正确检测到的攻击数目(right attack)。

Wc 为被误判(wrong)为攻击的连接数目。

这些数值在训练数据中均可准确计算，则适应度函数可以设计为

$$f(Xi)=\frac{Ra}{Ta}-\frac{Wc}{Tc} \tag{1}$$

$$F(Xi)=f(Xi)-\sum_{i=1}^{n}f(Xi)\Big/n \tag{2}$$

式(1)代表某个个体的检测情况值，分布在区间[-1, 1]中，其中，Ra/Ta表示检出率(该值越高越好)，则 1-Ra/Ta表示漏检率，而Wc/Tc表示误检率(该值则越低越好)。考虑到在一个处理器中包含有多个个体，因此，可综合考虑所有个体的检测值，首先求出它们的平均检测值，然后用某个个体检测值减去该平均值，得到该个体检测值和平均值间的差值$F(Xi)$，即式(2)作为本文的适应度函数。该值越高表明其适应能力越强，就可保留做进一步的遗传操作[3]，如果该值较低则被淘汰。在经过一定代数的进化后利用该函数就可以找到优秀个体。

3　系统实现主要模块分析

整个系统的实现共分 5 大模块，特征提取模块负责提取和建立入侵特征，入侵特征建立后需要用粗粒度模型遗传算法模块进行优化，以获得较优秀的特征个体并在处理器之间实现交流，每个处理器都可以利用这些含有入侵特征的个体进行入侵的检测，由检测模块实现这一功能。而如果要在实际的网络系统中进行实验的话，则还需要数据包采集模块来实现网络数据包的采集，以及发现入侵时让告警模块来实现入侵警告信息并记录下入侵数据以便让系统做出进一步的反应。

系统模型如图 2 所示。

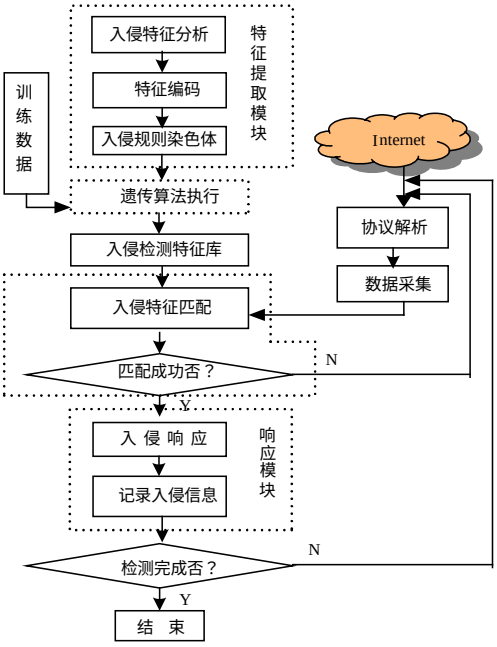

图 2　基于粗粒度模型遗传算法的网络入侵检测系统模型

3.1　特征提取模块

根据现有的入侵信息在网络协议中的具体体现和网络协议本身的结构特性容易被黑客所利用的特征信息，提取出这些特征并进行二进制编码，便于输送到算法模块中进行下一步的操作。

3.1.1　特征分析和提取

网络协议本身的结构缺陷常被入侵者所利用，如IP报文中的某些保留地址可以被攻击者用来产生恶意的数据包，如果协议中源IP地址=目的IP地址，则可以认为发生了Land攻击；TCP数据包中SYN和FIN标记同时被设置可绕过防火墙、路由器等。因此，可考虑将诸如IP源地址、IP目的地址、源端口、目的端口、IP报文标识域、TCP源端口、TCP目的端口、TCP确认号、ICMP的报文类型和报文说明、UDP校验和等最为重要的和攻击时最明显改变的特征[4]值提取出来作为入侵特征信息。

3.1.2　特征编码

编码采用二进制编码方式，即将上一步骤中所提取各协议的入侵特征信息，经过一定的组合和调配构成一个类染色体，并进一步转化为二进制(长度在 50 位左右)形式。如果提取过程中有不确定信息，也即某位既可表示为 1，也可表示为 0，那么可以用通配符“ # ”表示，因此，须先做如下规定：

染色体属性值用三元组{0, 1, #}来表示。

如果染色体某位为 0，则表示该位不会包含入侵信息，反之如果该位为 1，则表示在此可能是个入侵信号，而为#则表示 2 种情况都有可能，在后面的匹配过程中则 0 和 1 都可以与该位匹配。

3.2　粗粒度模型遗传算法模块

3.2.1　选择

按照随机的概率选择出已经初始化了的种群送往各个处理器，并保持各处理器中的种群数目一致，不至于增加某个处理器的负担，成为整个并行系统的瓶颈。

3.2.2　交叉

简单的交叉操作可分 2 步进行：(1)对处理器中个体进行随机的配对；(2)在配对的个体中随机地选择交叉点，对该点

前后的 3 个个体部分结构进行互换，并生成 2 个新个体，以增加算法的搜索面。用交叉概率*Pc*来控制交叉操作的频率，通常取值为 0.2～0.8 左右，该值太大虽可增强算法开辟新搜索领域的能力，但容易破坏适应度高的个体模式[5]；太小又不利于产生新的个体，使算法显得较为迟钝。

3.2.3　变异

针对单个个体进行，随机地选择个体的某一位进行变异(即取反操作：原来的 1 变为 0，原来的 0 变为 1)。该操作通过变异概率来控制。因其主要是维持群体的多样性，属于辅助性的操作，而且为了不使群体中的重要的基因丢失，取值通常在 0.001～0.01 之间。太大的话可能破坏很多好的模式；太小又不易产生新的个体，容易形成早熟现象。

3.2.4　适应度计算

经过选择、交叉和变异操作后，可以用训练数据(本文用 KDDCup99 数据)来进行染色体训练，然后统计正确检测到的攻击数和被误判为攻击的连接数目，首先计算出处理器中所有个体的检测情况值，然后调用设计的适应度函数即可得到每个个体的适应度，以便有选择性递进行淘汰和迁移。

3.2.5　迁移(采用一传一迁移策略)

根据处理器中个体的适应度函数值的情况决定迁移给相邻处理器的染色体的数目(send-best)。考虑到不增加系统的通信处理开销，该值可取在 2～8 之间，同时淘汰适应度最差的个体。然后选取适当的迁移间隔(send-time)，保证各处理器在经过适当的遗传代数后交换自己较优秀的个体，该值可取为 3～5，即当进化代数是该值的倍数时各处理器就交换一次自己的优秀个体。该值太大不利于优秀基因在各个处理器之间传播，也不利于最优解的寻找；而太小的话又使处理器通信过于频繁，使算法运行速度较慢[6]。

3.3　检测模块

遗传算法运行完成后，系统的入侵特征库就已经建立完成，该模块主要完成入侵信息的检测。将测试数据经过过虑并做简单处理后送往处理器进行匹配，如果测试数据中的某段或某位与特征库中的数据段(位)匹配，那么可判断该协议数据中包含有入侵信息，可通过告警模块作出进一步的反映；如果不匹配，表明没有入侵信息，则循环将后面的数据继续送往处理器进行检测。

4　主要算法实现

```
Begin
协议特征分析并编码；
构建入侵规则染色体；
int g=0；                    //进化代数 g 初始化为 0
while (g<=100)
{                            //进化代数最大值为 100 随机的选择染色
//体送往各个处理器；
double Pc=0.5；   //设定交叉概率进行交叉操作；
double Pe=0.005；  //  设定变异概率进行变异操作；
调用训练数据进行染色体训练；
统计染色体检测情况数据；
调用适应度函数 F(Xi)计算适应度值；
int Sb=4；                   //设置最大迁移个数
int St=5；                   //设置迁移间隔
if g%St=0
{
各处理器将自己最好的 Sb 个个体迁移到相邻的处理器；
接受来自相邻处理器的最好个体；
}
else
 淘汰适应度最差的个体；
g=g+1；                      //迁移算法
}
将网卡设置为混杂(promiscuous)模式，以便接收网络上的所有
//数据包；
利用 WinPcap 对流经网络的数据包进行捕捉并进行解析；
送往检测模块进行匹配检测过程；
If   匹配成功
{
送入侵响应模块处理；
记录入侵信息；
}
else   循环送入检测数据进行匹配；
while(检测完成 )
结束并退出；
End
```

5　实验结果及分析

实验用的处理器系统为 Win XP, CPU 为赛扬 D 2.4 GHz，内存为 512 MB，硬盘为 40 GB, 10 MB/100 MB 自适应集成网卡。实验数据采用的是 KDD Cup 99 测试数据集(此实验只选取了其中的一部分数据来进行，约有 15 MB，共 14 000 多条记录)。

实验过程中初始种群中共有 14 条染色体，每条染色体编码后长度为 50 位，进化代数为 100 代，定交叉概率设为 0.5，变异概率为 0.005，最大迁移个数为 4，迁移间隔为 5(每进化 5 代便进行一次迁移)。实验数据见表 1(只列出部分常用协议)。

表 1　实验数据记录表

协议类型	测试数据数目	入侵条目数	检测到的数目	检出率/(%)	误检数目	误检率/(%)
TCP	4 286	1 837	1 744	94.94	87	2.03
IP	3 758	1 593	1 476	92.66	52	1.38
ICMP	3 365	1 059	953	89.99	38	1.29
UDP	2 914	1 015	928	91.43	35	1.20

6　结束语

遗传算法在入侵检测系统中的应用属于较新的研究领域。通过把遗传算法引入到网络入侵检测系中，经过不断地对系统进行训练和测试，并逐步改进编码方式和算法，最终给出一种基于一传一迁移策略模型遗传算法的网络入侵检测系统(NIDS)。通过实验可以得出如下结论：

(1)检测效果上系统达到了预期目标，和其他的网络入侵检测系统相比，检测情况较好；

(2)高检出率(低漏检率)常会伴随高误检率出现；

(3)各种协议编码和组合方式的不同也会出现检测效果上的差异。

由此可见，通过引入粗粒度模型遗传算法到网络入侵检测系统，不会占用通信系统太多的带宽，且该方法解决了网络流量和检测性能之间的矛盾。下一步的工作将考虑进一步扩充协议类型，改进编码方法，并对适应度函数和系统算法再次进行优化设计，争取在误检率和漏检率之间找到一个较好的平衡。

参考文献

[1] 唐正军. 网络入侵检测系统的设计与实现[M]. 北京: 电子工业出版社, 2002.

（下转第 171 页）