DOI:10.13657/j.cnki.gxkxyxb.2013.04.011

广西科学院学报
Journal of Guangxi Academy of Sciences

2013,29(4):259～261
Vol. 29, No. 4 November 2013

基于遗传算法的入侵检测技术研究

The Study of Intrusion Detection based on Genetic Algorithm

贾 邓,杨 颖
JIA Deng, YANG Ying

(广西大学计算机与电子信息学院,广西南宁 530004)
(School of Computer and Electronical Information, Guangxi University, Nanning, Guangxi, 530004, China)

摘要:为了提高入侵检测在网络安全应用领域的效率以及准确度,对当前各类入侵检测技术方法的优缺点进行了比较,重点对基于遗传算法的入侵检测技术进行了分析,最终通过较好的基因编码以及改进的适应度构造函数,提高了基于遗传算法的入侵检测的效率及准确度,并使用仿真实验验证了改进后方法的有效性。

关键词:入侵检测 遗传算法 适应度函数

中图法分类号:TP393.028 **文献标识码**:A **文章编号**:1002-7378(2013)04-0259-03

Abstract: To improve the efficiency and accuracy of intrusion detection technology, this paper analyses the advantages and shortcomings of current existing intrusion detection technologies, especially the intrusion detection technologies based on genetic algorithm. Finally, the efficiency and accuracy of intrusion detection based on genetic algorithm was enhanced by improving encoding and fitness construction function, simultaneously the simulation results showed affectivity of the algorithm.

Key words: intrusion detection, genetic algorithm, fitness function

入侵检测系统(Intrusion Detection System,简称IDS)技术是目前国内外网络安全领域的重要技术和研究方向[1]。防火墙只能进行被动的防御,而入侵检测是积极主动的安全防护检测技术,可以对网络环境中的内部攻击、外部攻击和误操作进行实时保护[2]。作为被动防御防火墙之后的又一道安全防线[3],IDS可在很大程度上弥补防火墙的被动,而且能在不牺牲网络性能的前提下对网络进行全面检测。IDS的入侵检测方法对系统中没有访问权限的异常现象、活动进行跟踪、审查、识别并进行检验。IDS能识别出系统是否被入侵,从而在第一时间做出相应的反应(自动切断网络连接并记录时间、报警等),并发出警告提醒系统管理员采取应对措施。

当前主要的入侵检测技术有2种分类方式:数据来源和检测方法。按照数据来源进行分类,入侵检测技术可以分为基于主机的入侵检测与基于网络的入侵检测[4]。按照检侧方法进行分类,入侵检测技术可以分为基于误用的入侵检测和基于异常的入侵检测[5]。最早应用到入侵检测中的方法是统计分析方法,其系统随时更新记录用户行为,并通过概率统计分析方法检测用户的使用情况来查看该用户是否合法。但是该方法属于事后分析,无法及时进行入侵检测,而且其统计分析阀值[6]比较难确定。神经网络方法也被应用到入侵检测系统中,IDS使用该方法来学习用户日常行为[7],并模拟用户行为。神经网络方法允许模糊数据,但当神经网络面对大容量入侵行为时,占用大量资源的学习、解释的能力会在很大程度上降低。数据挖掘技术应用到入侵检测中的方法是系统通过从审计数据中学习用户行为来判断访问异常,主要应用死记硬背、监督学习、归纳学习、类比学习等方法[7]。

进化计算技术是对生物进化过程进行的一种仿生学研究。进化计算有5种不同的算法,分别为:遗

收稿日期:2013-04-12
修回日期:2013-08-14
作者简介:贾 邓(1982-),男,硕士研究生,主要从事数据挖掘、计算机网络研究。

传算法、进化策略、进化规划、分类器系统和遗传规划。其中，遗传算法是基于生物的自然选择，在计算机上进行模拟生物进化的一种寻优搜索算法，是应用比较广泛的全局优化方法。在遗传算法应用中，对实际问题进行抽象建模编码是遗传算法的主要启动步骤。适应度函数决定了算法是否能够使问题求解得到快速收敛。因此，问题的初始编码和适应度函数是遗传算法在入侵检测系统中应用的关键问题。

1 基于遗传算法入侵检测技术的关键问题

1.1 遗传算法

在基于遗传算法的入侵检测系统中，首先需要划定入侵行为和正常行为的特征库，通过应用遗传算法来搜索整个度量空间，实现特征库及其属性的最优化，以适应度函数评价的方式来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。

生物通过染色体之间的交叉和染色体本身的变异来完成种群的进化，遗传算法中最优解的搜索过程是通过模仿生物的进化过程。首先使用遗传算子作初始组成种群；然后进行选择运算、交叉运算和变异运算3种操作得到新一代种群；最后利用适应度函数来评估该种群是否是最优解，若不是，则继续进行选择变异操作，直到找到最优解结束。遗传算法的运行过程如下：

(1)对目标问题进行建模，并进行编码；

(2)随机生成初始化种群 $P(0)$；

(3)计算当前种群 $P(t)$ 中的每个个体 x_i 的适应度 $f(x_i)$；

(4) 对种群 $P(t)$ 进行选择运算，生成中间种群 $P'(t)$；

(5) 对中间种群 $P'(t)$ 应用交叉和变异算子，生成新种群 $P(t+1)$；

(6)$t=t+1$，并判断是否满足终止条件，如果不满足终止条件，算法跳回到第(3) 步继续运行；如果满足，算法结束，并输出结果。

文献[8]描述了二分类入侵检测技术和多分类入侵检测技术，将条件熵遗传算法的特征抽取与SVM训练模型进行联合优化的入侵检测技术。在采用遗传算法进行特征抽取时，改进交叉变异算子的设计，结合分类正确率和条件熵对特征子集进行评估。

1.2 检测器编码方式

随着遗传算法研究的不断发展，编码方式得到不断完善充实。最初的编码方式用二进制数据进行编码，类似于生物染色体结构，个体的每一基因位都是二进制数0或1进行编码。

整数编码：个体中的每一个基因都用一定范围内(如 $1\sim n$) 的整数表示，如整数规划中的TSP问题，可以采用整数编码进行求解。

浮点数编码：个体中的每个基因都用一定取值范围内的实数来表示。采用实数编码能提高算法的求解效率和解的质量，特别是在变量较多的情况下。

混合编码：实际问题中决策变量往往多种多样，既有开关状态的变量，也有实数型变量。在这种情况下，为了减少编码长度，采用混合编码方式是最好的选择。

在基于遗传算法的入侵检测训练算法中，通常以TCP/IP数据包为研究对象，其中染色体是由数据包中的版本、首部长度、服务类型、总长度、标识、标志、片偏移、寿命、协议、首部校验和、源地址、目的地址、源端口、目的端口、序号、确认序号、数据偏移、标志、窗口、校验和、紧急指针组成，并设定各自的影响权值 v_j $(0\leqslant j\leqslant 21)$。利用整数编码方式或者利用实数编码方式对获取的数据包进行初始化。利用各个基因的权值与实际值乘积的累加和确定入侵区间，通过一系列的选择变异等操作，应用相应的适应度函数，就可以最终挑选出最佳染色体。

1.3 适应度函数

为了获得最佳染色体，需要构造一个适应度函数。假设训练数据的总数为 N，根据各个基因的权值 $v_j(0\leqslant j\leqslant 21)$ 与实际值 $V_j(0\leqslant j\leqslant 21)$ 的乘积累加和 $C_i=\prod v_j^i\times V_j^i(0\leqslant i\leqslant N,0\leqslant j\leqslant 21)$ 会分布在不同的区间，设每个区间的长度为 L_i $(0\leqslant i\leqslant N)$，分布在每个区间的数据包数量为 $A_i(0\leqslant i\leqslant N)$，那么入侵的数据就应该分布在某个区间内。选择入侵区间关系到入侵检测系统的正确率、漏报率等很多重要指标，在正常情况下，入侵的数据一般都占少数，因此，选择区间中包含数据包最少的为异常区间，即阈值。根据这个原理，适应度函数可以构造为

$$\text{Fitness}=1-\frac{A_i}{N}\ (0\leqslant i\leqslant N)。$$

在不安全环境中，不主动与外界交互，接收到的数据中入侵的数据却能占大多数，此时，适应度函数可以构造为

$$\text{Fitness}=\frac{A_i}{N}\ (0\leqslant i\leqslant N)。$$

2　仿真分析

2.1　实验数据

实验数据来源于麻省理工学院林肯实验室提供的KDDCUP1999数据集。林肯实验室搭建了一个模拟美国军方局域网的实验环境,通过TCP Dump软件监听了9个星期而得到的网络连接数据,KDDCUP1999数据集的训练集和测试集两部分中包含了监听到的大量网络连接信息。每条连接信息提取基本特征集、内容特征集、流量特征集和主机流量特征集4类数据信息。

2.2　实验设计与参数设置

为了减少实验成本和提高实验的效率,初始检测器种群未采取随机生成的方法,而是利用KDDCUP1999数据集中已知的攻击数据的特征,进行排列重组生成初始检测器种群,大小为400。

选择算子采用杰出者选择策略,并采用父子混合选择,从而丰富检测器种群。变异算子使用均匀点变异,变异概率$P=0.4$。进化代数为$G=200$。

2.3　结果分析

实验在Matlab 7.0环境中运行。通过对各类数据集(Probe、DOS、U2R、R2L)的测试集进行实验,主要从检测率和误报率2个方面来分析算法的实用性,结果见图1。

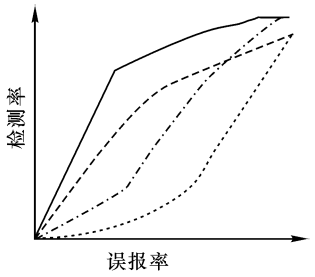

图1　本文检测算法与其他检测算法ROC曲线的比较
——:GA;－－－－:Cluster;－・－・－:K-NN;……:SVN。

由仿真结果可以看出,对特征进行维数约减和空间变换后,正确检测率和误报率都能取得满意的结果。使用ROC曲线来描述检测率与误报率,并与基于Cluster、基于K-NN和基于SVN的检测技术对比,可以看出,基于遗传算法的入侵检测技术在正确检测率和误报率的综合性能上都优于其他算法。

3　结束语

将遗传算法应用于入侵检测系统,可以提高入侵检测效益,实现全面、快速高效的检测出网络入侵。后续工作将致力于入侵检测系统探测器特征集的研究。

参考文献:

[1]　盛思源,战守义,石耀斌. 网络安全技术的研究和发展[J]. 系统仿真学报,2001,13:419-422.

[2]　Zhong Cheng,Li Na. Incremental clustering algorithm for intrusion detection using clonal selection[C]//Proceedings of 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. Los Alamitos,US: IEEE Computer Society Press,2008,1:326-331.

[3]　Zhong Cheng,Mi Ai zhong,Yang Feng. Intrusion detection using multiple classifiers fusion and clustering analysis[C]//Proceedings of the International Conference on Information Computing and Automation,World Scientific,2008,3:1181-1183.

[4]　李信满,赵大哲,赵宏,等. 基于应用的高速网络入侵检测系统研究[J]. 通信学报,2002,23(9):1-7.

[5]　朱文涛,李津生,洪佩琳. 基于路由器代理的分布式湮没检测系统[J]. 计算机学报,2003,26(11):1585-1590.

[6]　郭汉,曹元大. 入侵检测中攻击模式的挖掘[J]. 北京理工大学学报,2003,23(2):212-214.

[7]　唐正军,李建华. 入侵检测技术[M]. 北京:清华大学出版社,2004.

[8]　魏宇新. 网络入侵检测系统关键技术研究[D]. 北京:北京邮电大学,2008.

(责任编辑:尹　闯)