改进遗传算法优化的 BP 神经网络入侵检测研究

周贵旺,孙 敏

(山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要：入侵检测是一种主动的安全防护技术,能够对网络内部和外部的攻击进行防御。基于神经网络的入侵检测是常用的智能检测方法，其中 BP 神经网络是比较常用的神经网络模型。针对 BP 神经网络算法易陷入局部极值和收敛速度慢等问题,将神经网络与遗传算法相结合,用改进的遗传算法优化 BP 神经网络权值。

关键词：入侵检测;BP 神经网络;遗传算法

中图分类号：TP393.08 **文献标识码**：A **文章编号**：1674-7720(2010)21-0065-04

Improved genetic algorithm to optimize BP neural network for intrusion detection

ZHOU Gui-Wang, SUN Min

(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

Abstract: Intrusion detection is a proactive security technology to defend the attack from internal and external of the network. Intrusion detection based on neural network is commonly used intelligence test method, in which BP neural network is commonly used neural network model. In this paper, BP neural network algorithm easily trapped into local minima and slow convergence problems, the neural network and genetic algorithm combining genetic algorithm with the improved BP neural network weights. Actual results show that the method can effectively improve the training accuracy and the detection rate, shorten the training time.

Key words: intrusion detection; BP neural network; genetic algorithm

随着 Internet 的飞速发展，网络应用的种类不断增加,网络入侵手段也不断更新,网络安全问题已经成为信息社会所面临的最重要的问题之一。入侵检测作为网络与信息安全技术中新的研究领域和成果之一,必将对保障网络与信息的安全起到重要作用。目前国内外 IDS (Intrusion Detection System) 研究中所涉及的一些技术和方法主要有:基于神经网络的入侵检测技术、基于专家系统的入侵检测技术、基于 Agents 的入侵检测技术和基于模型推理的入侵检测技术等。其中,基于神经网络的入侵检测技术是近几年网络安全问题研究的热点之一[1]。

在基于神经网络的入侵检测技术中，反向传播 BP (Back Propagation) 神经网络是比较常用的神经网络模型。BP 算法本质上属于梯度下降算法,虽然具有自学习能力、寻优精确等特点,但如果初始连接权值取值不当,就会导致网络振荡、收敛速度慢、容易陷入局部极值等问题。针对 BP 算法存在的这些缺点，目前有一种解决方法是用具有全局搜索能力的遗传算法 GA (Genetic Algorithm)优化 BP 神经网络,将其应用于入侵检测。

普通 GA 的适应度函数不灵敏，其选择方法易产生随机误差,通用性较差,影响算法的性能。本文对 GA 的适应度函数和选择方法进行改进，用其优化 BP 神经网络,并应用在入侵检测中。

1 BP 神经网络及遗传算法简介

1.1 BP 神经网络简介

BP 神经网络模型是由一个输入层、一个或多个隐含层和一个输出层组成的一种多层前馈型网络，并用 BP 算法进行训练,是一种有导师的学习方法,利用梯度下降法对权值进行修正。在实际应用和研究中通常一个隐含层就能满足要求。

BP 算法的过程可以分为两个阶段。第一阶段是由输入层开始逐层计算各层神经元的输入和输出,直到输

出层为止。第二阶段是由输出层开始逐层计算各层神经元的输出误差，并根据误差梯度下降原则来调节各层的连接权值和节点阈值，使修改后的网络的最终输出能接近期望值[2]。如果一次训练以后还达不到精度要求，可以重复训练，直到满足训练精度为止。BP 神经网络模型流程图如图 1 所示[3]。

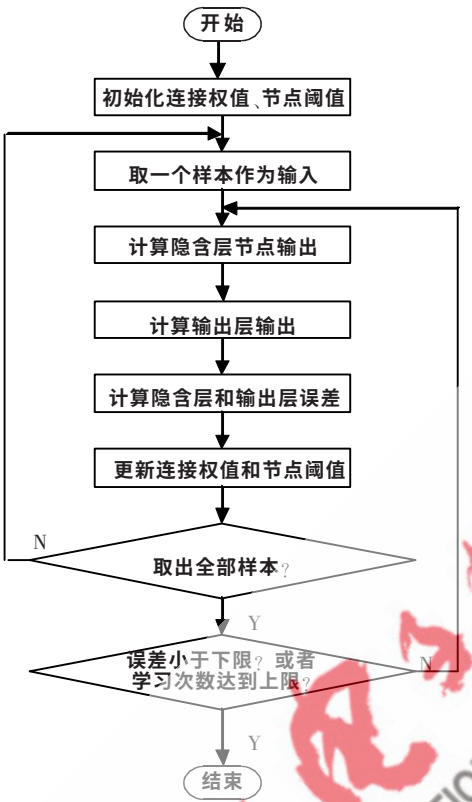

图 1 BP 算法流程图

1.2 遗传算法简介

遗传算法是由 DARWIN 的生物进化论和 MENDEL 的遗传理论发展而来的一种高效的全局搜索算法，它模拟自然选择和遗传中发生的繁殖、交配和突变现象，从初始种群出发，根据适应度函数计算出的适应度函数值，通过选择、交叉和变异这 3 个操作，产生新的更适应环境的个体(问题的解)，使群体进化到搜索空间中越来越接近问题的最优解的区域。这样一代一代不断进化，最后收敛到最适应环境的个体上，即求得问题的最优解。图 2 为遗传算法示意图。

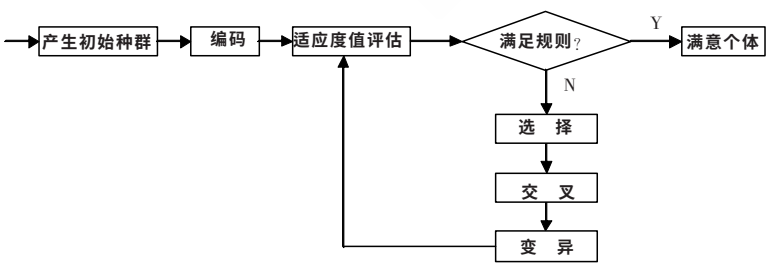

图 2 遗传算法示意图

在遗传算法求解问题时，个体的优劣程度是由适应度函数值的大小来判定的。通常对高于平均适应度值的个体做交叉，而对低于平均适应度值的个体进行变异，从而一代一代地提高群体的平均适应度值。对于同一种群而言，采用不同的适应度函数，平均适应度值就会不同，优于平均适应度值的个体数目也不同，即求解问题的能力有差别。由此可见，适应度函数在遗传算法求解问题的过程中起着至关重要的作用。

遗传算法中的选择、交叉和变异这 3 个操作是实现优胜劣汰进化的关键过程。理想情况下，如果每次选择操作选取的都是最适合解决问题的那些解，那么最后得到的解即为最优解。因而，选择方法在遗传算法优化过程中是非常重要的。常见的选择方法有轮盘赌选择法、锦标赛选择法和随机遍历选择法，这些方法各有特点，按照收敛速度由快到慢依次为：锦标赛选择法、随机遍历选择法和轮盘赌选择法。然而，锦标赛选择法和随机遍历选择法在选择的时候由于具有很大的随机性，容易产生随机误差，均不易找到全局最优解而陷入局部最优解。而轮盘赌选择法虽能找到全局最优解，但是收敛速度很慢。因此，遗传算法有必要进行改进。

2 遗传算法的改进

2.1 适应度函数

根据设计适应度函数的规范性、合理性、计算量和通用性等规则，本文采用参考文献[4]中设计的适应度函数：

$$Fit(f(x))=\begin{cases}1-0.5\times\left[\left|\dfrac{f(x)-b}{a}\right|\right]^{\alpha}, & |f(x)-b|<a\\ \dfrac{1}{1+\left[\left|\dfrac{f(x)-b}{a}\right|\right]^{\beta}}, & |f(x)-b|\geqslant a\end{cases}\tag{1}$$

在理想情况下，b 的值是 $f_{\min}(x)=y^*$，当适应度值为 0.5 时，α 是 $f(x)$ 到 $f_{\min}(x)$ 的距离。考虑到适应度函数的不同应用场合，本文将 β 值取为 2，将 α 值分别取为 1、1.5、0.5。

式(1)中的 a 和 b 随遗传算法的下一代进化而不断地修正。b 可取当前第 i 代中的最小值，即 $b=f(x)\times i$，而 a 用公式：

$$a=\max\times\left[0.5,\frac{|f_{\max}(i)-f_{\max}(i)^*|}{30}\right]\tag{2}$$

2.2 基于种群交流的选择方法

基于种群交流的选择方法就是利用两种或两种以上的选择方法，综合各种选择方法的优点，既能保证找到全局最优解，又能保证以一个相对较快的速度收敛，所以其性能通常较好。本文采用的基于种群交流的选择方法综合运用轮盘赌选择法和锦标赛选择法。

以群体 A 和群体 B 为例详细说明基于种群交流选择方法的基本思想[5](如图 3 所示)。群体 A 和群体 B 是两个不同的种群。在进化中，群体 A 第一代中的 a_{11} 与群体 B 第一代中的 b_{12} 产生的后代

a_{11}'和a_{12}'进入到群体 A 中的第二代；群体 B 第一代中的b_{11}与群体 A 第一代中a_{12}产生的后代b_{11}'和b_{12}'进入到群体 B 中的第二代。而群体 A 第一代中剩余的个体进行轮盘赌选择，群体 B 第一代中剩余的个体进行锦标赛选择。以后的每一代都按照上述的方式进化，直至达到最大进化代数 N[5]。

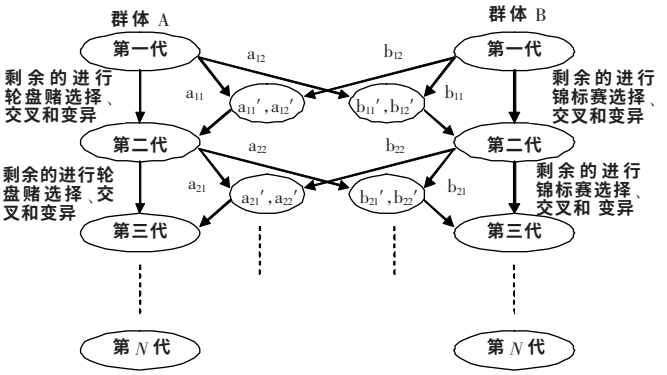

图 3 基于种群交流的选择方法

从图 3 中可以看出，在遗传算法进化过程中如果对每一个种群的每一代都进行种群间交流，当最大进化代数 N 很大时会严重影响算法的执行效率，而且效果也不一定好。所以在该方法具体应用过程中，以一定的概率（称为种群交流概率）随机对某两个种群进化过程中的某些代进行交流。这样既能发挥种群交流的优点，又有助于提高算法的效率。

3 基于改进遗传算法优化的 BP 神经网络原理

用遗传算法来优化神经网络可以分为三种：优化神经网络权值、优化神经网络结构和优化神经网络学习规则。因为神经网络的全部思想都体现在权值上，所以采用遗传算法优化神经网络权值，能够更好地提高神经网络的整体性能。用遗传算法优化神经网络权值的主要思想是改善神经网络的初始权值和节点阈值[6]。

本文用改进的遗传算法优化 BP 神经网络权值的主要思想是：初始化神经网络权值后，首先用 BP 神经网络进行训练，如果能满足精度要求就结束；如果不能满足精度，再用改进的遗传算法对 BP 神经网络权值进行优化，在解空间中找出一个较好、较小的搜索空间；然后，用 BP 算法在这个较小的解空间中搜索出最优解[7]。该算法的主要步骤如下：

(1)初始化连接权值和节点阈值，先用 BP 神经网络进行训练。若满足训练精度，则停止训练；否则，将这些初始权阈值初始化为一个初始种群，用改进的遗传算法进行优化。

(2)编码，在遗传算法中，编码影响着算法的性能和种群的多样性。二进制编码和实数编码相比较而言，二进制编码比实数编码搜索能力更强，而实数编码在变异操作上能更好地保持种群的多样性。本文采用这两种编码相结合的方式[8]，对网络结构采用二进制编码，对权阈值范围、学习速率和动量因子采用实数编码[9]。

(3) 用适应度函数计算出各初始种群对应的适应度函数值。

(4)选择，采用基于种群交流的选择方法。

(5)交叉，将选择后得到的新的群体按照预先确定的交叉率用均匀交叉的方式进行交叉。

(6)变异，依据预先给定的变异率进行变异操作。

(7)重复进行步骤(4)、(5)、(6)，直至满足达到最大进化代数后结束。

(8) 将得到的权值再次用 BP 神经网络训练判断是否满足精度要求。若满足，则算法结束；否则，继续对该权值进行训练，直至达到精度要求为止。算法流程图如图 4 所示。

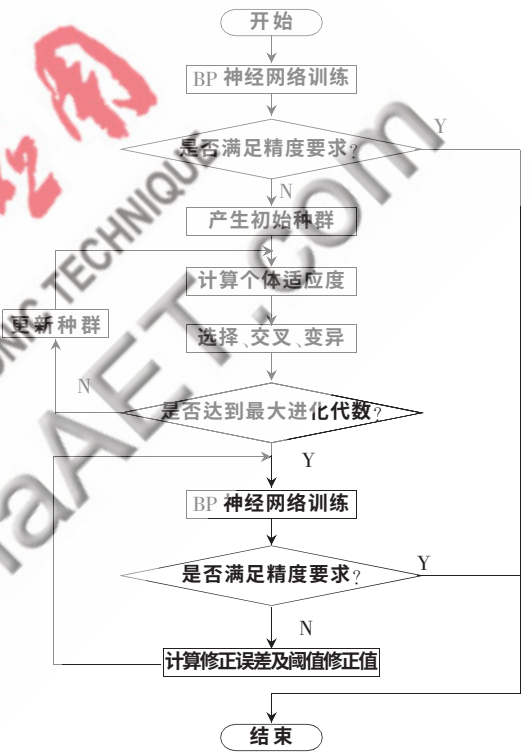

图 4 遗传神经网络算法流程图

4 入侵检测 MATLAB 仿真

4.1 数据获取

将本文提出的入侵检测方法用 Matlab 7.6 进行仿真，验证其性能。样本数据采用美国麻省理工学院林肯实验室提供的网络攻击评估数据[10]。数据集中的每条记录包含了 41 个特征量，根据实验环境和研究需要，选择每条记录的持续时间、协议类型、服务类型、目的端发送到源端的字节数、连接状态等 10 个特征作为研究对象，共选取 1 200 条记录。为了使神经网络能够处理非数值型数据，对数据特征中数值特征和非数值特征统一进行数值编码，并进行归一化处理。

4.2 参数设置

整个神经网络输入层节点为 10，隐含层节点为 15，输出层为 1；隐含层传递函数为 tansig，输出层传递函数

为 logsig，训练函数为 traingda，目标精度为 0.001，最大训练周期为 1 000；遗传算法初始种群大小为 200，最大进化代数为 200，选择概率为 0.9，交叉率为 0.8，变异率为 0.09，种群交流概率为 0.6。

4.3 仿真结果比较

将所选取的数据应用到遗传算法优化的神经网络和改进的遗传算法优化的神经网络，并用 Matlab 7.6 对两种神经网络分别进行入侵检测仿真，所得结果如图 5 和图 6 所示。

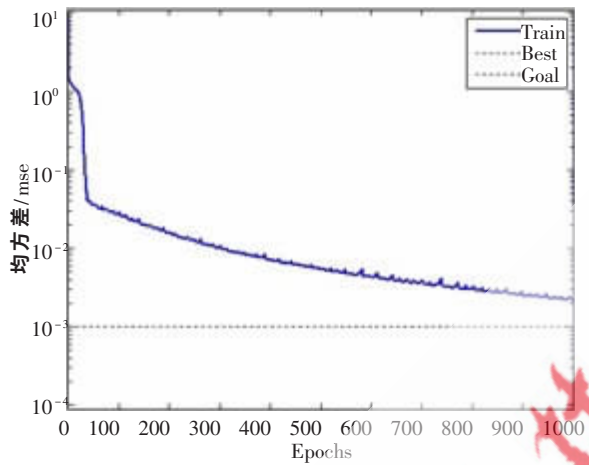

图 5 基于遗传算法优化的 BP 神经网络入侵检测仿真结果

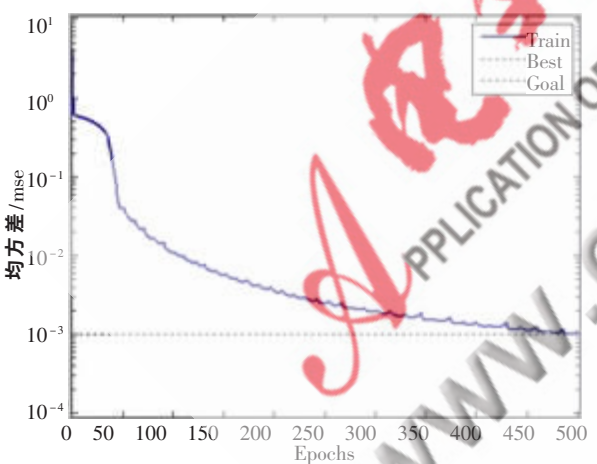

图 6 基于改进遗传算法优化的 BP 神经网络入侵检测的仿真结果

表 1 为对图 5 和图 6 的仿真结果进行的比较和分析，可以看出：基于改进遗传算法优化的 BP 神经网络入侵检测取得较好的效果。采用基于遗传算法优化的 BP 神经网络入侵检测在最初收敛很快，在 50 Epochs 后明显变缓，在 1 000 Epochs 时仍未收敛，所用的时间比较长；而采用改进遗传算法优化的 BP 神经网络入侵检测在开始 50 Epochs 里收敛比较慢，但在 50 Epochs 后收敛速度明显加快，在 502 Epochs 时收敛。在检测效果上，由表 1 可以看出，基于遗传算法优化的 BP 神经网络入侵检测的检测率为 89.67%，而改进的遗传算法优化的 BP 神经网络入侵检测的检测率为 96.58%，其误差、漏报率和误报率也明显提高。由此可见，采用改进遗传算法优化 BP 神经网络的入侵检测，训练精度、漏报率、误报率和检测率都明显提高，时间也比前者缩短了一半以上，性能较好。

表 1 仿真结果比较表

算法	误差	漏报率/%	误报率/%	检测率/%
GA-BP	0.00 099 995	7.35	2.98	89.67
改进的 GA-BP	0.00 075 236	2.56	0.86	96.58

本文主要是对遗传算法的适应度函数和选择方法进行改进，用改进的遗传算法优化 BP 神经网络，并将其应用在入侵检测中。有效克服了 BP 神经网络容易陷入局部极值和收敛速度慢等问题，增强了全局搜索能力，提高了训练精度、漏报率、误报率和检测率，缩短了训练时间，从而提升入侵检测的性能。

参考文献

[1] 王永全.入侵检测系统(IDS)的研究现状和展望[J].通信技术，2008，41(11)：139-146.

[2] 栾庆林，卢辉斌.基于神经网络与遗传算法的入侵检测研究[J].电子测量技术，2008，31(5)：70-74.

[3] 戴天虹.基于遗传神经网络的入侵检测研究[J].中国安全科学学报，2006，16(2)：103-108.

[4] 刘英.遗传算法中适应长函数的研究[J].兰州工业高等专科学校学报，2006，13(3)：1-4.

[5] 魏全新，刘贤锋，黄锵，等.遗传算法选择方法的比较分析[J].通信和计算机，2008，5(8)：61-65.

[6] LAM H K. Tuning of the structure and parameters of neural network using an improved genetic algorithm [C]. Industrial Electronics Society, Denver, CO, USA: IECON'01, 2001.

[7] 栾庆林，卢辉斌.改进遗传算法在神经网络权值优化中的应用研究[J].遥测遥控，2008，29(1)：51-54.

[8] YEN G G, LU H. Hierarchical genetic algorithm based neural network design [C]. 2000 IEEE Symposium on Combinations of Evolutionamy Computation and Neural Networks, 2000: 168 -175.

[9] 马海峰，宋井峰，岳新.遗传算法优化的混合神经网络入侵检测系统[J].通信技术，2009，42(9)：106-108.

[10] http://www.ll.mit.edu./IST/ideval/data/2000/LLS-DDOS-2.0.0.html.2001-06-01.

(收稿日期：2010-07-03)

作者简介：

周贵旺，男，1984 年生，硕士研究生，主要研究方向：网络安全。

孙敏，女，1968 年生，副教授，硕士生导师，主要研究方向：网络安全。