

CIPS 中基于改进 GANN 的入侵检测模型

张正光, 李国宁, 陈 璐

(兰州交通大学自动化与电气工程学院, 兰州 730070)

摘 要: 应用在计算机集成过程系统(CIPS)网络中的入侵检测系统误报率和漏报率较高。针对该问题, 利用遗传算法的全局寻优能力和神经网络对于非线性映射的强大逼近能力, 提出具有自学习和自适应能力、基于遗传算法神经网络(GANN)的入侵检测模型, 包括数据采集模块、数据预处理模块、神经网络分析模块和入侵报警模块 4 个部分。为克服遗传算法易早熟、搜索迟钝的缺点, 对 GANN 的适应度值调整方式进行改进, 对遗传算法的参数设定进行优化, 并采用改进的遗传算法优化收敛速度慢、易陷入极值的 BP 神经网络。仿真实验结果表明, 该模型使系统的检测率提高至 97.11%。

关键词: 遗传算法神经网络; BP 神经网络; 入侵检测; 计算机集成过程系统; 主成分分析

Intrusion Detection Model Based on Improved Genetic Algorithm Neural Network in Computer Integrated Process System

ZHANG Zheng-guang, LI Guo-ning, CHEN Lu

(School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

【Abstract】 In order to solve the problems of high false alarm rate and fail rate in intrusion detection system of Computer Integrated Process System(CIPS) network, this paper takes advantage that Genetic Algorithm(GA) possesses overall optimization seeking ability and neural network has formidable approaching ability to the non-linear mapping to propose an intrusion detection model based on Genetic Algorithm Neural Network(GANN) with self-learning and adaptive capacity, which includes data collection module, data preprocessing module, neural network analysis module and intrusion alarm module. To overcome the shortcomings that GA is easy to fall into the extreme value and searches slowly, it improves the adjusting method of GANN fitness value and optimizes the parameter settings of GA. The improved GA is used to optimize BP neural network. Simulation results show that the model makes the detection rate of the system enhance to 97.11%.

【Key words】 Genetic Algorithm Neural Network(GANN); BP neural network; intrusion detection; Computer Integrated Process System(CIPS); Principal Component Analysis(PCA)

DOI: 10.3969/j.issn.1000-3428.2013.04.036

1 概述

随着我国铁路的高速发展, 其网络安全问题越来越突出。计算机集成过程系统(Computer Integrated Process System, CIPS)是编组站自动化控制系统和综合管理信息系统的综合集成。编组站 CIPS 必须实现实时、基于规则和管理集中的入侵检测, 从而在网络边界和网络内部检查和响应来自外部的攻击和可疑行为, 同时可中断内部多事者和外部黑客的非授权使用、误用和滥用^[1-2]。在编组站综合集成自动化系统网络中应用入侵检测技术能够有效地避免网络攻击, 但是传统的入侵检测系统存在如下问题: (1)具有自学习、自适应能力的入侵检测系统还远未成熟; (2)现有

的入侵检测系统误报率、漏报率较高, 对未知的网络攻击检测能力差。

目前, 主要的研究成果有: 文献[3]提出了一个完整的基于数据挖掘的入侵检测框架, 在入侵检测领域数据挖掘方面做出了开创性的工作; 文献[4]提出了基于人工免疫的入侵检测模型; 文献[5]提出基于标准遗传算法的入侵检测, 但是标准遗传算法^[6-7]有易早熟、搜索迟钝等缺点, 主要表现在开始随机选取个体时有一定的盲目性, 最终无法保持种群的多样性而产生早熟现象。

针对上述问题, 本文提出基于改进的遗传算法神经网络(Genetic Algorithm Neural Network, GANN)的入侵检测模

基金项目: 铁道部科技研究开发计划基金资助重点项目(2011X008-D)

作者简介: 张正光(1984—), 男, 硕士研究生, 主研方向: 网络安全, 交通运输自动化与控制; 李国宁, 副教授; 陈 璐, 硕士研究生

收稿日期: 2012-04-25 **修回日期:** 2012-06-20 **E-mail:** zzguang1225@163.com

型, 将标准的遗传算法进行改进, 用线性调整法对适应度值进行调整, 由此提高入侵的检测率。

2 CIPS 网络中入侵检测模型设计

本文的入侵检测模型包括数据采集模块、预处理模块、神经网络分析模块、入侵报警模块, 其结构如图 1 所示。

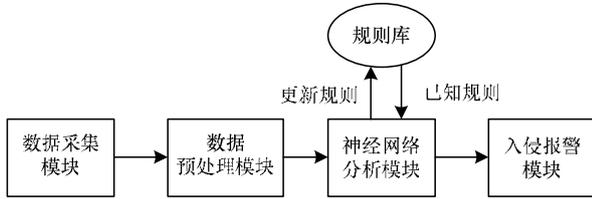


图 1 入侵检测模型结构

(1) 数据采集模块

由于编组站 CIPS 系统目前主要是 Windows 系统平台, 因此利用 Winpcap 数据包采集程序, 通过调用其中的 Packet.dll 和 Wpcap.dll 中 2 个动态连接库提供的 API 函数, 将网卡上收到的 CIPS 网络数据包收集起来^[8]。

(2) 数据预处理模块

该模块将原始数据转为神经网络可识别的数据, 对原始数据的处理可分为 3 个部分: 协议分析, 数据包特征选取, 数据处理。由于数据间数值的差异较大, 应该对原始数据进行归一化处理, 设输入数据为 $S = \{x_1, x_2, \dots, x_n\}$, 则归一化处理步骤如下:

1) 计算平均值

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

2) 计算标准差

$$\sigma(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

3) 归一化

$$x'_i = \frac{x_i - \bar{x}}{\sigma(x)}$$

(3) 神经网络分析模块

遗传神经网络模块学习导入的训练样本, 用遗传算法调整各层的连接权值, 不断对其进行调整, 经过多次训练后, 神经网络相对稳定, 入侵行为就能被较准确地识别出来。

(4) 入侵报警模块

为了方便 CIPS 管理人员进行网络监管, 入侵报警模块的功能是进行系统报警提示, 经过神经网络的判断, 如有入侵行为发生, 则产生提示报警, 否则, 系统正常。

3 改进的 GANN

遗传算法虽然有宏观搜索能力强、全局寻优等优点, 但是在应用中易早熟、搜索迟钝^[9-11]。改进的遗传算法通常的方法有: (1) 编码; (2) 确定适应度函数; (3) 设定遗传算法

自身参数, 包括种群大小 n 、交叉率 p_c 和变异率 p_m 。本文将对适应度值的调整方式进行改进, 并优化遗传算法的参数设定。

3.1 编码

采用实数编码的方式对 BP 神经网络的初始权值进行编码, 操作方法是: 随机产生具有 N 个个体的初始种群 $x = \{x_1, x_2, \dots, x_n\}$, 任意 x_i 为 BP 神经网络的连接权值和阈值的初始值, 每个基因表示一个连接权值或阈值, 个体的长度 $n = i \times a + a \times b + a + b$, 其中, a 为隐层节点数; b 为输出层节点数; i 为输入层的节点数。

3.2 适应度函数选取

选取的适应度函数为:

$$fitness = A/E$$

$$E = \sum_{i=1}^N (y_i - t_i)^2$$

其中, $fitness$ 为适应度函数; A 为常数; E 为所有样本的误差平方和; N 为样本的个数; y_i 为网络的实际输出值; t_i 为网络的期望值。

适应度值一般有 3 种调整方法: 窗口法, 函数归一化法, 线性调整法。线性调整法是个有效的调整方法。设调整后个体适应度值为 F , 原个体适应度值为 f , 则 $F = af + b$; 系数 a 、 b 可通过多种方法选取。在任何情况下均要求 F_{avg} 与 f_{avg} 相等, 应满足的条件为:

$$\begin{cases} F_{avg} = f_{avg} \\ f_{max} = c_{mult} F_{avg} \end{cases}$$

其中, c_{mult} 是经验值, 是最佳种群所要求的期望副本数, 正常条件下的线性调整方法如图 2 所示, 对 50~200 规模的种群来说, 取值范围为 1.2~2。一些个体的适应度值远远小于平均适应度值和最大适应度值, 而往往平均适应度值和最大适应度值又非常接近, 这是线性调整在遗传算法后期可能产生的问题, 本文采用的解决方法如图 3 所示, 将 c_{mult} 原始适值伸展成负值, 当没有合适的 c_{mult} 时, 仍保持 $F_{avg} = f_{avg}$, 而将 f_{min} 映射到 $F_{min} = 0$ 。

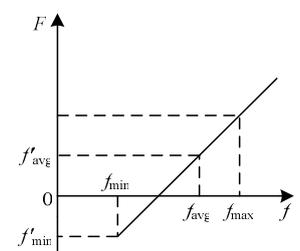
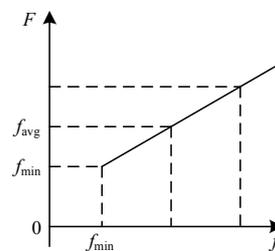


图 2 正常条件下的线性调整 图 3 特殊条件下的线性调整

3.3 遗传操作

遗传操作包括以下三部分:

(1) 选择: 遗传算法中最常用的选择方法是适应度比例

法，或叫“轮盘赌选择”策略，在该方法中，每个个体的选择概率和适应度成正比。设种群的大小为 n ，第 i 个个体的适应度为 f_i ，其被选择的概率为：

$$P_{si} = f_i / \sum_{i=1}^N f_i$$

(2)交叉：采用两点交叉，设位串的字符长度为 l ，在 $[1, l-1]$ 内，随机地选取一个整数值 k 作为交叉点。将 2 个配对串从第 k 位右边部分的所有字符进行交换，从而生成 2 个新的位串。

(3)变异：采用均匀变异，确定一个较小的区间 $[-A_i, A_i]$ ， $i=1, 2, \dots, n$ ，当 v_k 变异时，随机均匀地在 $[-A_i, A_i]$ 中取一个 y ，令 $v_k = v_k + y$ 。其中， A_i 称为变异域，一般取区间 $[a_i, b_i]$ 长度的百分比，如 $A_i = (a_i, b_i) \times 0.1$ 。

该算法步骤如下：

步骤 1 设定初始种群为 n ，随机产生一个初始种群。

步骤 2 计算每个个体的适应度。

步骤 3 排序后适应度低的个体被淘汰。

步骤 4 以 p_c 为交叉率进行交叉操作产生新的个体，直接复制没有进行交叉操作的个体。

步骤 5 以 p_m 为变异率变异产生新个体。

步骤 6 把新个体插入到种群中，计算新个体的适应度。

步骤 7 计算神经网络的误差平方和，如果达到期望值，转步骤 8；否则，转步骤 4，继续进行遗传操作。

步骤 8 用 BP 算法训练神经网络，直到期望的精度，结束训练。

算法操作流程如图 4 所示。

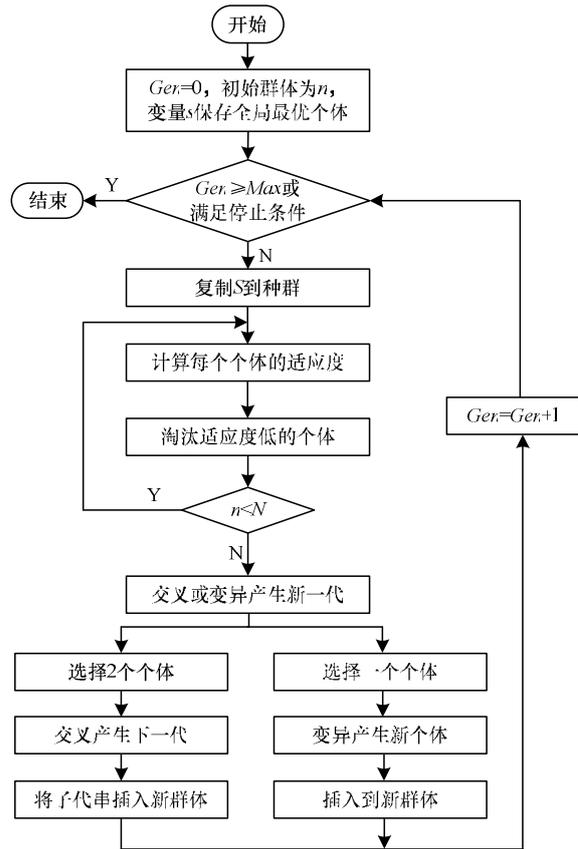


图 4 改进的遗传算法流程

4 遗传神经网络结构

本文所使用的神经网络分为输入层、隐含层和输出层。

(1)输入层和输出层的设计

输入层和输出层的神经元个数可以根据需要和数据表示方式确定。根据对预处理模块提取到的 41 维特征向量进行数据处理，经主成分分析法(Principal Component Analysis, PCA)降维后的主元变量是 7，可以设输入层神经元个数为 7，输出层神经元个数确定为 1，也可根据实际情况改变输出层神经元的个数。

(2)隐含层的设计

确定隐含层的神经元个数是个复杂的问题，不能仅用解析式获得该数目，需要经验和多次实验来确定。确定隐含层神经元的个数常用如下公式：

$$\sum_{i=0}^n C_{n_i}^i > k$$

其中， k 为样本数； n 为输入层神经元个数。

如果 $i > n$ ， $C_{n_i}^i = 0$ ，则 $n_1 = \sqrt{m+n+a}$ ，其中， m 为输出层神经元个数； n_1 为隐含层神经元个数， $n_1 = 1bn$ ； a 为 $[1, 10]$ 之间的常数。根据分析与计算，设计的 BP 神经网络为 7 个输入、1 个输出，包含 10 个隐含层单元，结构如图 5 所示。

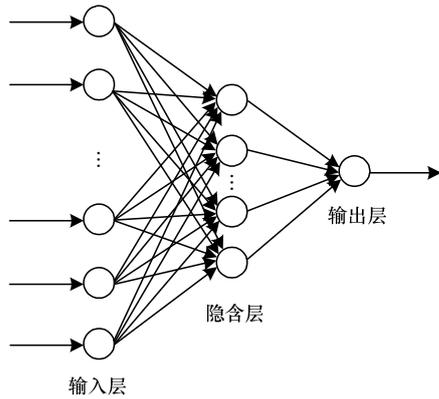


图 5 GABP 网络结构

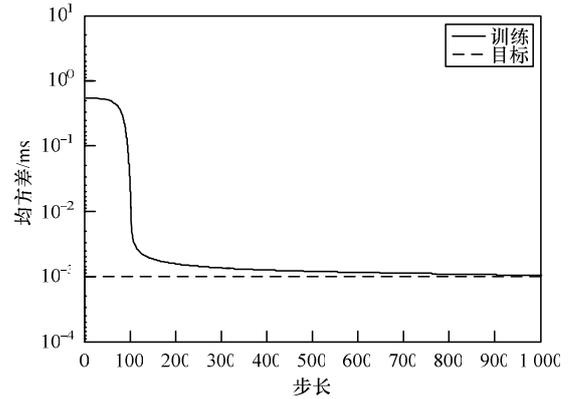


图 7 改进的 GANN 收敛结果

5 仿真实验与分析

5.1 实验参数设置

设遗传算法的初始种群大小为 200, 最大进化代数为 500, 选择概率为 0.9, 交叉率为 0.35, 变异率为 0.09。

在经典的 BP 算法中, 训练速率由经验确定, 训练速率越大, 收敛越快, 但训练速率过大, 将会引起振荡。因此, 在不导致振荡的前提下, 训练速率越大越好, 该值一般取 0.6~0.9。

最大训练次数设为 2 000 次, 超过 2 000 次时就停止。

最小误差设为 0.001, 当输出误差小于这个设定值时, 训练结束。

5.2 实验数据选取

本文采用的实验数据来自 MIT 的数据集 KDD CUP99, 这些数据已成为近年来评判入侵检测系统的标准数据^[12]。从 KDD CUP99 数据中随机抽取约 500 条样本数据作为系统的训练和测试数据。在 4 大类攻击(Probe、U2R、DoS 和 R2L)中抽取 15 种典型的攻击作为本实验的基础数据。

5.3 实验结果分析

实验利用 Matlab7.0 进行遗传神经网络的实验。为了验证改进的遗传算法对入侵检测的有效性, 本文分别用未改进的遗传神经网络与改进的遗传神经网络进行比较, 对比结果如图 6、图 7 所示。

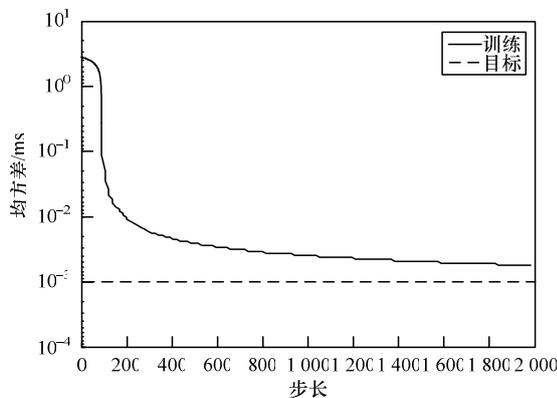


图 6 未改进的 GANN 收敛结果

从中可以看出, 基于未改进的遗传神经网络的入侵检测在最初 80 步的时候收敛速度很快, 但之后收敛速度明显放慢, 在 2 000 步时仍未收敛, 所用时间较长。而基于改进的遗传神经网络的入侵检测在 100 步时收敛速度较快, 在接近 956 步时已经达到收敛的效果。

由表 1 可见, 基于改进的遗传神经网络的入侵检测效果较好, 收敛速度快, 所用时间短, 检测率有明显的提高, 误报率和平均检测时间都有所下降, 所以, 本文入侵检测模型的检测率、误报率、漏报率、平均检测时间均最好。

表 1 3 种检测模型的性能比较

入侵检测模型	检测率/(%)	误报率/(%)	平均检测时间/s
BP	79.43	20.25	0.49
遗传 BP	93.23	6.32	0.43
改进的 GANN	97.11	1.23	0.31

6 结束语

本文将改进的遗传算法应用于 CIPS 网络入侵检测中, 建立基于 GANN 的入侵检测模型, 并进行了仿真实验, 结果表明其检测率有明显的提高, 因此, 可以实现对 CIPS 的有效管理与控制, 确保编组站的安全与高效生产。铁路编组站网络的安全防护有防火墙、入侵检测、杀毒工具等, 如何将入侵检测与其他防护措施进行相互配合, 进一步提高检测效率, 是下一步的研究内容。

参考文献

- [1] 胡 鹏, 吴振兴. CIPS 环境下的网络安全设计[J]. 铁路通信信号工程技术, 2007, 4(1): 50-53.
- [2] 余 荣. 探析编组站网络系统安全问题[J]. 铁路通信信号工程技术, 2009, 6(6): 26-28.
- [3] Lee W, Stolfo S J. A Data Mining Framework for Building Intrusion Detection Model[C]//Proc. of 1999 IEEE Symposium on Security and Privacy. Oakland, USA: IEEE Computer Society Press, 1999: 120-132.
- [4] Forrest S, Perelson A S, Allen L, et al. Self-nonsensical

- Discrimination in a Computer[C]//Proc. of IEEE Symposium on Research in Security and Privacy. Oakland, USA: IEEE Computer Society Press, 1994: 202-212.
- [5] Shon T, Seo J, Moon J. SVM Approach with a Genetic Algorithm for Network Intrusion Detection[C]//Proc. of the 20th International Symposium on Computer and Information Sciences. Berlin, Germany: Springer-Verlag, 2005: 224-233.
- [6] 易晓梅, 陈 波, 蔡家楣. 入侵检测的进化神经网络研究[J]. 计算机工程, 2009, 35(2): 208-213.
- [7] 刘衍鹏. 基于改进的遗传神经网络入侵检测系统的应用研究[D]. 重庆: 重庆大学, 2009.
- [8] 曹宏丽. 入侵检测技术在微机监测网络中的应用研究[D]. 兰州: 兰州交通大学, 2011.
- [9] 栾庆林, 卢辉斌. 自适应遗传算法优化神经网络的入侵检测研究[J]. 计算机工程与设计, 2008, 29(12): 3022-3024.
- [10] 狄文辉. 基于改进量子遗传算法的入侵检测特征选择[J]. 计算机测量与控制, 2011, 19(4): 813-815.
- [11] 胡明霞. 基于 BP 神经网络的入侵检测算法[J]. 计算机工程, 2012, 38(6): 148-150.
- [12] 郑洪英, 侯梅菊, 王 渝. 入侵检测中的快速特征选择方法[J]. 计算机工程, 2010, 36(6): 262-264.

编辑 张 帆