人工神经网络技术在系统流量异常检测模块中的应用

张艳萍，高忠新，于全勇
(牡丹江师范学院网络信息中心，黑龙江 牡丹江 157012)

摘 要：介绍人工神经网络技术，建立了人工神经网络的典型模型。应用BP算法的泛化功能,将输入输出样本进行训练,不断学习调整网络权值,使网络实现给定的输入输出映射关系,以达到检测流量异常的目的。

关键词：人工神经网络技术；流量检测
中图分类号：TP393.0 **文献标识码**：A

Application of neure network technology in the intelligent system flowing anomaly detection module

ZHANG Yan Ping, GAO Zhong Xin, YU Quan Yong
(Madanjiang Normal College Network Information Center, Mudanjiang 157012, China)

Abstract: In this paper, artificial neural network technology and the establishment of a typical artificial neural network model are introduced. Application of BP algorithm functions generalization, the input and output sample training and through learning and adjusting the value of the network, so the network is scheduled to map the relationship between input and output, and achieve the purpose of detecting abnormal flow.

Key words: neural network technology; flowing detection

1 人工神经网络原理及算法实现

1.1 人工神经网络的工作原理

人工神经网络首先要以一定的学习准则进行学习，然后才能工作。所以网络学习的准则是：如果网络作出错误的的判决，则通过网络的学习，可以减少下次犯同样错误的可能性。经过网络按学习方法进行若干次学习后，网络判断的正确率将大大提高。

连接机制结构的基本处理单元与神经生理学类比称为神经元。每个构造起网络的神经元模型模拟一个生物神经元，如图1所示。该神经元单元由多个输入信号(i=1, 2,...,n)和1个输出y组成。中间状态由输入信号的权和表示

$$y_j(t) = f(\sum_{i=1}^{n} w_{ji}x_i - \theta_j) \qquad (2\text{-}1)$$

而输出模型如图所示。

1.2 BP算法分析

采用BP算法(反向传播学习算法)网络模型分析网络异常。BP网络学习的主导思想是通过不断调整权值，使误差代价函数最小，标准的BP算法采用的是一阶梯度法，即最速下降法。

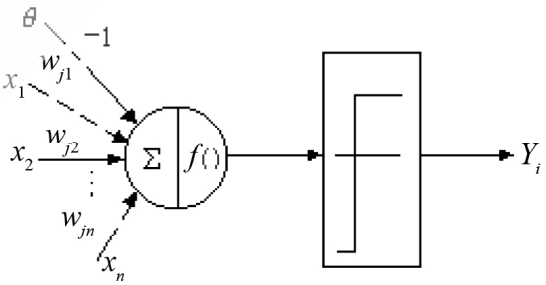

图1 神经元模开型

如果有M层网络，而第M层仅含输出节点，第一层为输入节点，则BP算法设计为：

第一步：选取初始权值W。

第二步：重复下述过程直至收敛：

(1) 对于k=1 ~ N

① 计算Oik、netjk和yk的值(正向过程)。

② 对各层从M ~ 2反向计算(反向过程)。

(2)对同一节点$j \in M$，计算δjk;

第三步：修正权值，$Wij=Wij-\mu\frac{\partial E}{\partial W_{ij}}$，$\mu>0$，其中$\frac{\partial E}{\partial W_{ij}}=\sum_{k}^{N}\frac{\partial E_k}{\partial W_{ij}}$。

从上述BP算法可以看出，BP模型使得一组样本的I/O问题变为一个非线性优化问题。

设计神经网络专家系统重点在于模型的构成和学习算法的选择。结构是根据所研究领域及要解决的问题确定的。通过对所研究问题的大量历史资料数据的分析及目前的神经网络理论发展水平建立合适的模型，并针对所选的模型采用相应的学习算法。在网络学习过程中，不断调整网络参数，直到输出结果满足要求。

1.3 BP神经网络在流量异常检测模块中的应用

流量异常检测模块的目的是在某个时间段内检测出某个子网某个端口的流量是否出现异常。通过对4个校园子网流量连续21天观察记录，利用BP神经网络进行网络流量训练后，继续记录网络流量数据，并输入到相应的BP神经网络中，利用训练后的权值矩阵进行计算，如果误差结果大于设定的最小误差，则认为此流量数据异常，并记录了此数据的子网、子网掩码、端口号，以供异常分析模块使用。

1.3.1 数据源的选取

对流量的监测可以有几种可能的数据源，根据对牡丹江师范学院网通出口进行连续21天的连续观察，通过对得到的数据分析，发现进出校园网的数据包有较强的规律性，同时包数的突增突减也能反应网络的流量异常，因此适合作为神经网络的数据源。而进出校园网的流量字节数变化较大，不适合进行报警。

1.3.2 样本数据的选定

如果希望能在某个时间段内检测出某个子网某个端口的流量是否出现异常，则需要构建这一时段、这一子网、这一端口的BP神经网络，因此共构建了子网数、端口数、时间段数这三者乘积的BP神经训练网络。分别以2小时、3小时、4小时为一个时段(INTERVAL)，并将每个时段划分为8个时区，每过一个时区就记录一次这个时区内的牡丹江师范学院实验楼、主楼、培训机房、电子阅览室这4个校园子网的数据包总量以及特定的6个端口数据包个数，每过一个时段就将这8个时区内累计的数据记录在特定文件中。这样一天内就可以得到24/INTERVAL(即时间段数)个文件，对于每个子网每个端口而言，一个文件就为它们提供了这一时段内的数据样本，其中包含了8个样本数据。连续监测21天，这样每个BP神经训练网络都有21个数据样本。每个样本中的8个数据，经过转换，成为BP神经网络输入层的8个输入数据，训练之后得到的权值矩阵就是某个时段内某个子网某个端口数据包流量的权值矩阵。

1.3.3 BP网络参数设定

(1)网络节点　网络输入层神经元节点数就是系统的特征因子(自变量)个数，输出层神经元节点数就是系统目标个数，隐层节点根据经验选取。在系统训练时，实际还要对不同的隐层节点数分别进行比较，确定出最合理的网络结构。所以设定了一个三层的BP神经网络，输入层有8个节点，输出层有1个节点，隐含层有两层，第一个隐含层包含30个神经元，第二层包含8个节点。

(2)初始权值的确定。初始权值是不应完全相等的一组值。已经证明，即便确定存在一组互不相等的使系统误差更小的权值，如果所设W_{ji}的的初始值彼此相等，它们将在学习过程中始终保持相等。因此，在程序中设计了一个随机发生器程序，产生一组-0.5 ~ +0.5的随机数，作为网络的初始权值。

(3)最小训练速率。在经典的BP算法中，训练速率由经验确定，训练速率越大，权重变化越大，收敛越快。但训练速率过大，会引起系统的振荡。因此，训练速率在不导致振荡的前提下，越大越好。该值一般取0.9。

(4)动态参数。动态系数的选择也是经验性的，一般取0.6 ~0.8。试验中取值0.7。

(5)允许误差。这个误差是由试验判断而来，为了取得较低的漏报率和误报率，对于不同时段，设定不同的允许误差，取值在0.01~0.02之间。

(6)迭代次数。一般取1 000次。由于神经网络计算并不能保证在各种参数配置下迭代结果收敛，当迭代结果不收敛时，允许最大的迭代次数。经过试验判定，选取了400次。

(7)Sigmoid参数。该参数调整神经元激励函数形式，一般取0.9~1.0之间。试验中选取0.9。

1.3.4 流程描述

变量描述：

(1)描述子网数据结构

```
struct Sub
{
unsigned long subnetip;          // 子网 ip
unsigned long submask;           // 子网掩码
}*sub;
```

(2)网络数据包计数数据结构

```
struct Cal_Value
{
struct Sub sub;                          // 子网情况描述
unsigned long  t_pnum[ZONE][PORT]; // 记录特定时间
                                   段内端口数据包数量的数组
} *cal_val;
```

1.3.5 训练流程

(1)数据收集。利用libnids提供的函数接口监听网络

数据，关注的是数据包头的信息。当源或目的地址与需要检测的校园子网 IP 地址相同时，再查看其端口是否是特定的端口(21，23，25，53，80，110，8 080)，是则相应的端口数据包计数加 1，否则记入最后一个计数变量(即此子网数据包总量计数值)。每过一个时区，将计数值记入到相应 cal_val 动态数组中；每过一个时段，创建一新文件，将 cal_val 数组中的数据写入此时段的文件中，文件名记入到一个数组 filename[ ]中，供以后的 BP 网络训练使用。当达到训练天数时，创建 BP 算法训练线程，终止数据收集。

(2)样本数据输入。利用数据收集阶段的 filename[ ]数组，对于每一个 BP 神经网络，选出各自的样本文件，将样本 8 个数据除以 1000 000(使 BP 神经网络输入值在(-1,1)之间)输入到神经网络。

(3)设置输出期望值为 1，设定 BP 网络拓扑、阈值。

(4)利用 BP 算法训练样本数据，经过正向传播、反向训练之后,得到权值矩阵，记录到相应的权值文件中，为检测流量异常做好准备。

1.3.6 异常检测流程

(1)数据收集与数据输入。BP 神经网络训练完成之后，仍按照训练阶段的方式收集数据，不同的是当记录完一个时段的一个数据文件后，立即启动检测异常线程，将文件中的数据经同样的转换，输入到相应的 BP 网络中，进行异常检测。

(2)BP 神经网络载入相应的权值文件，经过正向传播计算后，将计算结果与期望输出值进行误差计算，判断是否大过设定的阈值。若大于阈值，则认为出现异常，由此进入异常处理阶段。BP 神经网络计算线程结束，主线程继续收集数据。

(3)异常处理阶段。发现异常后，记录异常的时段、子网 IP 地址与掩码、子网端口号。而后启动流量异常分析模块，收集此子网该端口的所有数据包，应用数据挖掘技术进行分析，找出攻击特征。

2 实验结果分析及实验意义

2.1 实验环境和目的

试验在双 CPU2.4GHz，主存为 4GB 的戴尔机架服务器上进行，操作系统为 Redhat 9.0，硬盘为 146GB SCSI；网络为牡丹江师范学院校园网。实验主要是通过对样本数据进行多次训练，确定合适的阈值、时间段，使异常检测的漏报率、误报率达到相对较小。

2.2 实验数据

由于实验所得数据量较大，这里只选取一些典型数据的进行说明。在 16:00～21:00 之间，牡丹江师范学院电子阅览室的 80 端口网络流量较大，有较好的说明性。采用不同时段的 12 个神经网络的输出数据(即校验数据)来选定时间段、阈值。经过神经网络计算后，统计了各个阈值的误报率和漏报率。

2.3 结果分析

在 16:00～21:00 之间，实验以 2 小时、3 小时、4 小时为一个时段，选取不同的阈值，训练此时段 80 端口的 BP 神经网络。试验过程中，由于使用攻击工具，造成短时间内流量增大，以此确定阈值的最大数值；并且在校园网网络防火墙和 snort 这样的基于规则的网络入侵检测系统的帮助下，通过对误报的判定，确定阈值的最小值。由此在 3 个时段里，各选取了 3 个阈值，收集了 12 天的网络流量数据。这些样本数据里包含了异常数据和神经网络误报、漏报的数据。试验结果分别如表 1、表 2、表 3 所示。表中，W 为误报率，L 为漏报率，T 为正确检测。

由表可以看出，在同一个时段内，当阈值取值较小时，误报率较高，而漏报率较低；当取值较高时，误报率较低，而漏报率较高。如表 1 当阈值取值为 0.01 时误报率较小到达了 33.3%，而漏报率只有 8.3%；当阈值取值为 0.015 时，漏报率为 33.3%，误报率为 8.3%。相比较而言，当漏报率和误报率大致相当的时候，就能够取得相对较好的检测效果，

表1 4小时为一时段的试验数据

	样本1	样本2	样本3	样本4	样本5	样本6	样本7	样本8	样本9	样本10	样本11	样本12	误报率/%	漏报率/%
0.01	W	T	L	W	T	T	T	T	T	T	W	W	33.3	8.3
0.012	W	T	L	T	T	T	T	L	L	T	T	W	16.7	25
0.015	T	T	L	T	T	T	T	L	L	L	T	W	8.3	33.3

表2 以3小时为一时段的试验数据

	样本1’	样本2’	样本3’	样本4’	样本5’	样本6’	样本7’	样本8’	样本9’	样本10’	样本11’	样本12’	误报率/%	漏报率/%
0.015	T	W	T	W	T	L	W	W	T	T	T	W	41.7	8.3
0.018	T	T	T	W	T	L	T	W	L	T	T	T	16.7	8.3
0.020	T	T	L	T	T	L	T	W	L	L	T	T	8.3	33.3

表3 以2小时为一时段的试验数据

	样本1''	样本2''	样本3''	样本4''	样本5''	样本6''	样本7''	样本8''	样本9''	样本10''	样本11''	样本12''	误报率/%	漏报率/%
0.018	W	T	T	W	T	T	W	T	L	T	W	W	41.7	8.3
0.020	T	T	L	W	T	T	W	T	L	T	W	T	33.3	16.7
0.022	T	T	L	T	T	T	W	L	L	L	T	T	8.3	33.3

既能检测到绝大部分异常的发生，还能减小误报。如表1所示，阈值取为0.012时，检测效果最好。由此原则确定了各个时段的阈值，它们依次是0.012，0.018，0 .02。

纵向比较表1、表2、表3，对于同一阈值而言，时间段越小，它的误报率就越高；时间段越大，它的漏报率就越高。以表1、表2中的阈值0.015为举例，当时段大小为4小时，它的误报率为8.3%；而当时段大小为3小时，它的误报率为41.7%；表2、表3阈值为0.2的漏报率在时段大小为3小时，漏报率为33.3%，而在时段大小为2小时，误报率为16.7%。由此可以得出这样一个结论，为了得到较好的检测效果，时段的选择应该是该时段阈值的漏报率与误报率较为相当的时段。比较表中的0.012、0.018、0.02这3个阈值的误报率与漏报率，不难看出大小为3小时的时段，是应该选择的检测时段。

2.4 实验意义

实验通过人工神经网络技术实现流量异常检测，结合试验结果得出结论：通过连续21天观察记录网络流量和12天的校验数据的收集，确定了检测时段的大小为3小时，神经网络阈值为0.018,流量检测模块检测的漏报率、误报率达到相对较小，效果最好。这一检测方法对提高网络流量异常检测的准确性和检测效果具有普遍指导意义。

参考文献

[1] 王丽娜，董晓梅，于戈，等.基于进化神经网络的入侵检测方法[J].东北大学学报(自然科学版),2002(2).

[2] 吕昌国.基于BP算法的网格资源调度研究[D].哈尔滨:哈尔滨理工大学，2007.

[3] 周梦熊.基于实数编码遗传神经网络的入侵检测方法研究[D].哈尔滨:哈尔滨理工大学,2007.

(收稿日期：2008-12-10)

IT动态

控创双核3U CompactPCI® CPU板加入对LynxOS-SE实时操作系统的支持

在“中等健壮”级安全应用中表现出色

2009年1月20日， 全球领先的可即用（COTS）产品和定制化嵌入式解决方案供应商控创，和全球领先嵌入式软件供应商LynuxWorks™发表声明，在基于Intel®Core™2 Duo处理器的控创ITC-320 3U CompactPCI®CPU板中加入对LynuxWorks' LynxOS-SE 5.0的支持。控创ITC-320和LynxOS-SE 5.0的此番结合，可提供适合美国政府定义的“中等健壮”级安全应用所用到的计算机解决方案。

控创ITC-320系列属于3U CompactPCI® CPU板系列，该系列使用基于Intel®嵌入式架构，配备高性能芯片组，具备长期可用性。支持LynxOS-SE 5.0的CPU板所使用的处理器为1.5 GHz Intel®Core™2 Duo LV 。支持内存ECC校验，精心选择器件来保证产品的长生命周期以及为了优化散热而做的元器件布局设计都是产品设计思想的一部分，目的在于使控创ITC-320能满足安全应用市场日益变化的需求。软件支持上向后兼容PCI总线的PCI express端口（1个x4或4个x1），控创ITC-320是数据处理类应用的理想选择。

LynxOS-SE 5.0 BSP旨在满足安全、需实时处理的关键应用，使得国防人员可以快速受益于针对恶劣环境进行过特殊优化以及为实时系统提供了全套强大的、零宕机时间的软硬件。该操作系统兼容Linux ABI，全面支持POSIX ，提供了开放式API、通过单一级别操作系统级保护配置文件来实现的中级安全性保障和一系列高级网络功能。LynxOS-SE 5.0还对RAM进行了扩容，让开发人员能更好地利用当前硬件设计方案。该系统还拥有全面集成的、经过升级的GNU 工具，以及能提供流水线式开发和执行环境的强化Linux ABI。

（控创公司供稿）