神经网络在数字化校园安全检测中的应用①

张 领

(商丘医学高等专科学校 现代教育技术中心, 商丘 476100)

摘 要: 数据安全是数字化校园建设的重要问题, 快速准确检测数字化校园的安全性及存在的风险和漏洞, 成为急需解决的问题. 在改进 BP 算法基础上, 设计一个基于神经网络的数字化校园安全检测原型. 通过统计底层网络协议(TCP)的数据流量和信息数据包协议头的信息, 将信息预处理后送入已训练过的神经网络模块, 以此判断当前网络数据流量存在的攻击或扫描行为. 实现快速检测数字化校园存在的漏洞和安全隐患, 提前预防和减少数字化校园受到的攻击和破坏.

关键词: 数字化校园; 神经网络; BP 算法; 安全检测

Application of Neural Network in the Digital Campus Safety Testing

ZHANG Ling

(Shangqiu Medical College, Shangqiu 476100, China)

Abstract: Data security is an important problem in digital campus construction. It is urgent to find a way to detect the risk and vulnerabilities of digital campus security. Therefore, a digital campus etwork security detection prototype was designed based on the improved BP algorithm. The data flow and nformation coming from the underlying network protocols (TCP) will be collected and pre-handled. The trained neural network module will react to these data and information, which gives the clue to determine the existing attack or scanning behavior in current network data flow. Rapid detection of digital campus existing vulnerabilities and security risks can prevent and reduce the attack and destruction.

Key words: digital campus; neural network; BP algorithm; safety inspection

高等学校数字化校园建设正朝着数字化、信息化和网络化的目标迈进. 其建设水平代表着高校整体办学水平和现代化管理水平. 然而, 在数字化校园网络系统建设和运用过程中, 网络安全隐患日益突出, 诸如: 木马、蠕虫、恶意网络攻击、垃圾邮件以及采用高科技手段实施窃密和盗窃的事件在部分高等学校中时有发生. 确保校园网络信息资源的安全、可靠和完整传输是数字化校园网络安全的重中之重[1].

针对数字化校园的安全问题, 很多高校常用的检测方法有两种: 第一种是模式匹配法: 是常常被用于入侵检测技术中. 它是通过把收集到的信息与网络入侵和系统误用模式数据库中的已知信息进行比较. 从而对违背安全策略的行为进行发现. 模式匹配法可以显著地减少系统负担, 有较高的检测率和准确率. 第二种是专家系统法: 这个方法的思想是把安全专家的知识表示成规则知识库, 再用推理算法检测入侵, 主要是针对有特征的入侵行为[2]. 以上两种检测方式的不足之处是状态和转换动作是手工编码. 因而很难精确表达; 没有积极主动的一些安全防护方式相配合, 他的安全性能就是有残缺的. 因此, 开发一种主动防御、及时控制的的数字化校园安全检测系统就非常重要.

1 数字化校园安全检测的方法和技术

作为一种内部外部都可以动态检测的安全测试, 安全检测系统可以用来检测各种类型的入侵, 被看做

① 收稿时间:2015-05-18;收到修改稿时间:2015-06-15

防御系统的重要组成部分. 异常检测把当前活动概况与定义好的概况做对比, 若发现不一致, 就把当前活动认作是入侵行为. 把这种入侵行为写入活动概况中. 因此, 我们称异常检测又叫基于行为的检测. 这种检测行为时相对独立的行为, 通用性强[3].

2　人工神经网络和BP算法的改进

2.1 基于神经网络的入侵检测技术

BP 神经网络是应用最普遍的神经网络模型之一, 也称误差后项传播神经网络, 是一种由非线性变换单元组成的多层前馈网络, 一般由输入层、输出层、隐含层组成. 常见的三层 BP 模型如图 1 所示.

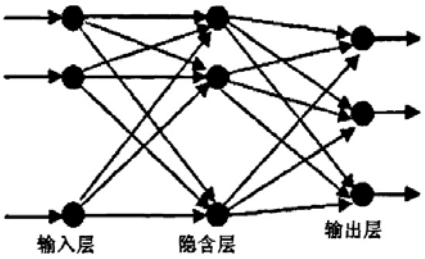

图 1　典型的 BP 模型结构

为了训练一个 BP 网络, 需要计算网络的输出和误差, 然后求得网络误差平方和. 当所训练矢量的误差平方和小于误差目标, 训练则停止, 否则在输出层计算误差变化, 并且采用反向传播学习规则来调整权值, 并重复此过程.

BP 算法的具体描述如下[4]:

输入: 训练样本, 学习效率, 多层前馈网络.

输出: 一个训练的、对样本分类的神经网络.

Stepl: 初始化神经网络的权值和偏值;

Step2: 当结束条件满足时, 继续, 否则转向 Step6;

Step3: 向前传播输入: 对隐藏层和输出的每个单元的误差, 计算每个隐层单元的误差;

Step4: 向后传播误差: 计算输出层每个单元的误差, 计算每个隐层单元的误差;

Step5: 更新权值和偏值, 转向 Step2;

Step6: 训练好的神经网络.

这种方法的优点在于具有良好的适应能力, 能够较好地处理异常数据, 更好地检测一些未出现过的新攻击类型, 降低网络复杂度, 提高识别能力. 但它的不足之处在于网络的拓扑结构存在学习能力的限制容量, 学习效率低, 收敛速度慢, 对于具体发生的事件类型缺乏明确的解释能力.

2.2 BP 算法的改进

现在用的最多的 BP 算法, 虽然与其他算法比有相对的优势, 但是也有收敛速度慢, 容易进入局部最小, 甚至造成网络无法正常运行等不可忽视的问题, 需要我们改进. 通过结合 Cauchy 训练逐步实现对 BP 算法的优化.

2.2.1 Cauchy 训练

如果将网络的训练看成是让网络寻找最低能量状态的过程, 取网络的目标函数为它的能量函数, 再定义一个初值较大的数为人工温度.

模拟退火组合优化法的基本思想是[5]: 随机地为系统选择一个初始状态$\{w_{ij}^{(p)}\}$, 在此初始状态下,给系统一个小的随机扰动$\Delta w_{ij}^{(p)}$, 计算系统的能量变化.

$$\Delta E = (\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\}) - E(\{w_{ij}^{(p)}\}) \quad (1)$$

若ΔE <0 则此扰动被接收. 如果ΔE > 0 则此扰动依据概率

$$\exp(-\frac{e}{kt}) \quad (2)$$

判断是否被接收. 如果此扰动被接受, 则系统从状态$\{w_{ij}^{(p)}\}$变换到状态$\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\}$, 否则系统的状态保持不变. 如此重复下去. 在这个过程中, 逐渐地降低温度T, 退火算法的流程图如图 2 所示.

模拟退火算法的具体描述及说明如下[5]:

1) 初始化各层的联接权矩阵 W; 定义人工温度T的初值;

2) 每一个温度T重复如下过程:

① 选取一个样本, 计算其输出与目标函数$E(\{w_{ij}^{(p)}\})$;

② 随机地从$\{w_{ij}^{(p)}\}$中选取一个$w_{ij}^{(p)}$;

③ 按一定地算法产生$w_{ij}^{(p)}$的一个调整量$\Delta w_{ij}^{(p)}$;

④ 按照$\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\}$重新计算相应的输出和目标函数$E(\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\})$

⑤ $\Delta E = (\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\}) - E(\{w_{ij}^{(p)}\})$;

⑥ $\Delta E > 0$ then

⑦ 用$\{w_{ij}^{(p)} + \Delta w_{ij}^{(p)}\}$代替 . . . ;

⑧ if 样本集中还有未被选用地样本 then 转(2.1);

3)降低温度 T;

4)如果 T 足够小, 则结束, 否则转(2).

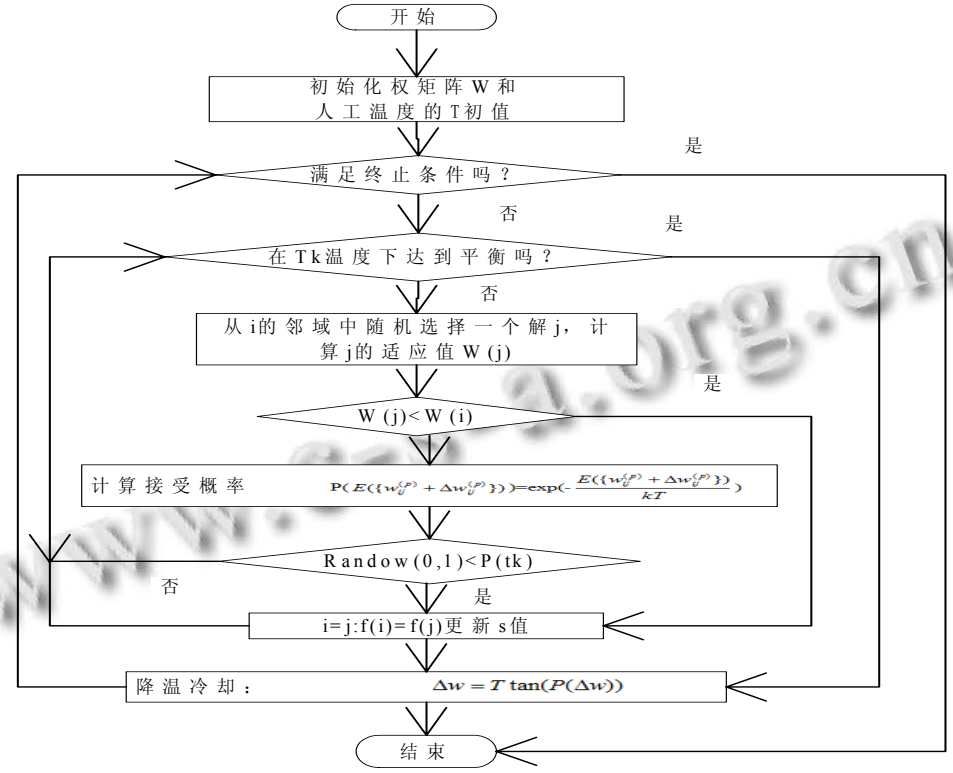

图 2 退火算法的流程图

3 数字化校园安全检测原型

数字化校园安全检测分为五个功能模块, 分别是数据捕获模块、数据预处理模块、神经网络训练模块、神经网络测试模块、响应报警模块.

3.1 数据捕获模块

网络数据的捕获通过 Libpcap 编程接口实现. 表 1 描述了接口函数及意义. 这种接口是网络数据包的标准捕获接口, 它效率高、独立性和移植性强. 使用具有快速的网络数据包过滤功能的 BPF(Berkeley Packet Filter)数据包捕获机制.

表 1 接口函数

char*pcap_lookupdev(char*errbuf)	将网络接口的名字函数形式返回
int pcap_iookupnet(register const char*device, register bpf_u_int32*netp, register bpf_u_int32*maskp, register char*errbuf)	网络地址和网络掩码的获取函数, device 是网络接口的名字, netp 是网络地址的存放指针, maskp 是网络掩码的存放指针.
int pcap_loop(pcap_t*P, int cnt, pcap_handler callback, u—char*user)	调用 callback 回调函数, 原型在回调函数 callback 中对捕获到的数据包进行预处理

3.2 数据预处理模块

数据预处理模块接收网络数据包, 这些数据包由数据捕获模块送过来. 接收后先处理这些数据包, 并将处理的结果转化为向量送入神经网络检测模块检测. 数据预处理模块处理的数据包分别针对三种协议: IP、TCP、ICMP. IP 协议的预处理主要提取针对基于 m 碎片的攻击特征; TCP 协议的预处理针对基于 TCP 协议网络扫描并对拒绝服务攻击特征提取.

TCP 协议预处理

对 TCP 协议解析函数中的 TCP 数据包分析和统计,

目的是为了能检测基于TCP协议的某些拒绝服务攻击和进行端口扫描. 在预处理中, 每一个 TCP 连接和半开 TCP 连接信息要记录下, 一个 struct TcpConnect 结构体变量对应一个连接或半开连接. 当前的连接状态有变量的 state 成员指示, 结构体变量通过它们的 pNext 指针成员连接成 SYN 链表, 在设定好的时间内超时的链表节点将被删除删除, 这样做一使得此链表总是反映最近一段时间的连接信息, 另一方面可以防止内存堆空间的耗尽. 对此链表节点一一访问并进行统计分析可得出对安全检测有价值的特征信息, 归一化处理这些信息即可做为标准特征向量送入神经网络检测模块进行检测. 其流程图如图 3 示.

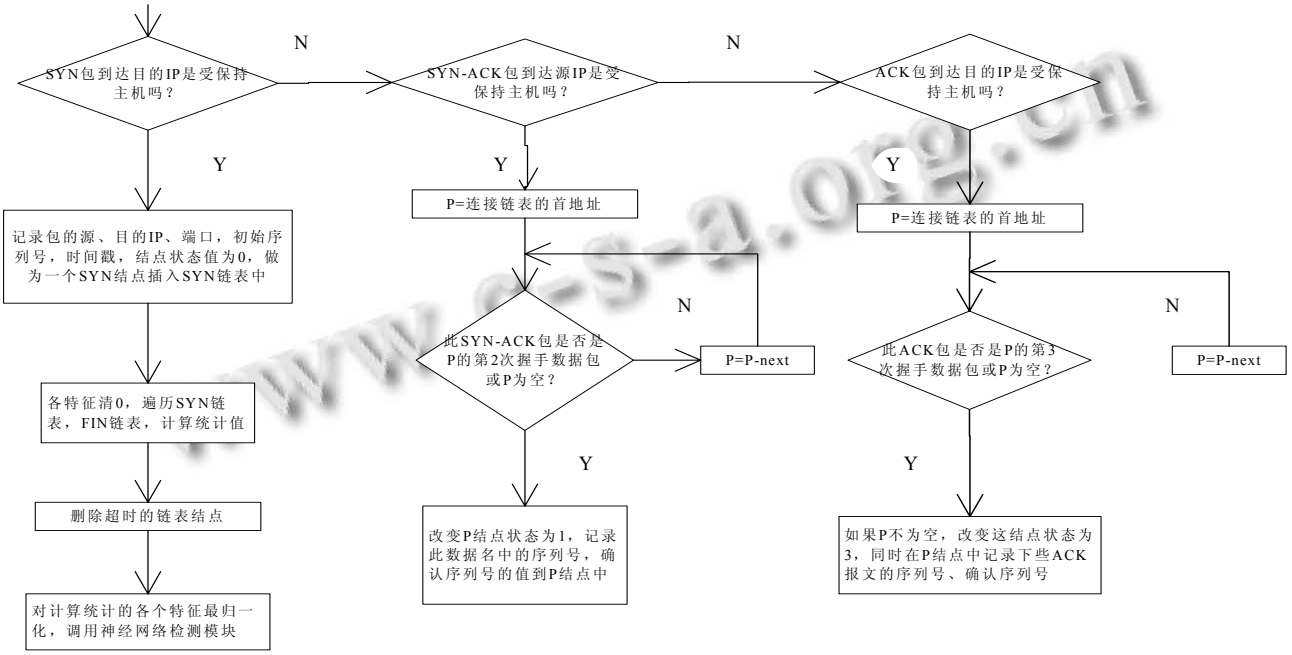

图 3　处理 TCP 连接信息流程图

3.3 神经网络训练模块

TCP 神经网络有以下几个输入特征向量: (1)在最近时间段与当前 SYN 报文目的端口、目的 IP 相同的 SYN 报文数; (2)在最近时间段与当前 SYN 报文目的端口、目的 IP 相同的 ACK 报文数; (3)在最近Δf 时间段与当前 SYN 报文目的端口、目的 IP 相同的 SYN 报文数/在最近出时间段与当前 SYN 报 文目的端口、目的 IP 相同的 ACK 报文数; (4)SYN 链表中不同目的端口数 p/SYN 链表长度 1 的数值; (5)在一个小的时间间隔Δf内指向被扫描主机不同端口 SYN 报文数; (6)在一个小的时间间隔Δf 内指向被扫描主机不同端口 FIN 报文数; (7-12)TCP 报文的 6 位标志位(URG、 ACK、PSH、RST、SYN、FIN). 在送入神经网络之前归一化处理特征, 归一化处理使神经网络的所有输入元素的取值范围在 0 到 1 之间, 归一化处理使神经网络训练模块、检测模块可以得到网络扫描的特征向量和标准的可描述攻击行为. 标准化的特征向量和对应的编码通过 TCP 协议的拒绝服务攻击和漏洞检查得到. 如表 2 示[6].

表 2　基于 TCP 协议的一些攻击和扫描的标准化特征向量

名称	标准化特征向量	神经网络期望输出的编码
TCP 连接耗尽拒绝服务攻击	1,1,0,0.00,0.05/0.10,0,0,0,0,0,0,1,0;	1,0,0,0,0,0;
SYN 风暴拒绝服务攻击	1,0,1,0.00/0.05/0.10,0,0,0,0,0,0,1,0;	0,1,0,0,0,0;
TCP 连接扫描	0,0,0/1,0.30/0.35/0.40/0.45/0.50/0.55/0.60,1,0, 0,0,0,01,0;	0,0,1,0,0,0;
TCP SYN 扫描	0,0,0/1,0.80/0.85/0.90/0.95/1.00,0,1,0,0,0,0,0,1	0,0,0,0,1,0;
FIN 扫描	0,0,0/1,0.00/0.05/0.10/…/0.95/1．00,0,1,0,0,0,0,0,1;	0,0,0,0,1,0;
正常的网络流量	有不同组合的 294 个标准输入特征向量	0,0,0,0,0,1;

SYN 风暴攻击和 TCP 连接耗尽攻击牵涉到 TCP 连接的第一次握手的SYN报文. 特征(1)的有大的数量, 结果 1 就是归一化后得到的; 第二个特征与特征(2)的取值有较大的差别, 原因在于 TCP 连接耗尽攻击涉及到整个 TCP 连接的三次握手, SYN 风暴攻击只涉及到前两次握手, 不涉及第三次握手的 ACK 报文. 1 和 0 特征(2)归一化之后得到的结果. 前两个特征决定第三个特征, 第三个特征归一化结果为 0 和 1. 特征(4)的取值是一个接近于 0 的小数, 原因在于涉及到不同目的端口数较少, SYN 链表的长度较大. 特征(5)(6)归一化之后取值都为 0; TCP 报文中的六个标志位, 不需再归一化处理, 它们只取 0 或 1. 这两种攻击, 当前有 SYN 报文到达是共同特征, 这些共同特征决定 SYN 标志位取 l,其它各标志位取 0. 通过表 5.1 中的基于 TCP 的网络扫描可以看到, 特征(4)(5)(6)和 TCP 六个标志位体现了他们的差别. 原型预处理模块处理之后得到特征的取值和标准化值, 输出到磁盘文件中[6].

3.4 神经网络检测模块

神经网络检测模块接受的特征向量是由预处理模块送来, 这些特征向量做为神经网络的输入向量, 是由神经网络的各层计算出结果, 然后把结果在输出层输出. 输出结果可能显示的是正常的数据流, 可能指示某一种网络攻击, 也可能是未知的结果即这个结果在神经网络训练时没有定义. 针对这种情况, 对于前两种情况, 系统管理员应记录下这时神经网络的输入向量和输出向量, 在神经网络进行训练时, 加入这两种情况. 对于后一种情况, 需要在进行分析.

3.5 响应报警模块

响应报警模块对接受神经网络输出向量, 根据编码的对应关系, 判断攻击行为. 发现对应的攻击. 就会有相应的消息打印在屏幕上, 并会有指示这种攻击的全局变量在系统中设置; 若没有找到对应攻击, 则不打印消息. 理论上讲训练神经网络时, 对输出向量的编码是 0、1 串. 但真正的输出不会严格的是 0、1 串, 所以约定, 数值大于 0.7 的按 1 对待, 小于 0.3 的按 0 对待.

4 结论

传统的数字化校园安全检测系统, 在攻击实际发生之前, 它们往往无法预先发出警报, 这样就有可能会造成系统进行安全检测时已经被攻击或入侵. 我们会始终处于备份还原、杀毒、制定新的安全策略等被动防御过程中. 数字化校园安全检测原型的设计中, 引入神经网络技术, 能够有效快速发现学校网络的漏洞和存在风险, 减少学校网络被攻击的可能性. 对学校数字化校园的安全性提供了保障.

参考文献

1 彭耘.入侵检测技术在数字化校园网中的分析与设计.电脑编程技巧与维护,2013,18:104–105.
2 张新淼.数字化校园安全应用防护策略研究与实践.软件导报,2014,12:151–152.
3 苏岩.一种面向数字化校园网络的安全设计方法.科技视界,2014,14:37–38.
4 张学锋,黄子辉.基于遗传算法的 BP 神经网络入侵检测系统在校园网中的应用.计算机与现代化,2010,18:18–19.
5 旷昀.BP 算法在高校计算机等级考试过级预测中的应用[学位论文].北京:北京工业大学,2007.
6 薛俊.基于神经网络的网络入侵检测技术的研究与实现[学位论文].南京:东南大学,2009.