

# 基于BP神经网络的病毒检测方法

郭晨, 梁家荣, 梁美莲

(广西大学计算机与信息工程学院, 南宁 530004)

**摘要:** 提出了一种基于BP神经网络的病毒检测方法, 该方法成功地把BP神经网络的理论引入计算机病毒的检测中。该方法比传统的病毒检测技术更有效地对系统信息和文件系统进行语法分析, 快速地诊断出被感染病毒以及病毒类型。

**关键词:** BP神经网络; 病毒检测; 计算机病毒; 训练样本

## Method of Virus Detection Based on BP Neural Networks

GUO Chen, LIANG Jiarong, LIANG Meilian

(College of Computer and Information Engineering, Guangxi University, Nanning 530004)

**[Abstract]** In this paper, a new method of computer virus detection on back-propagation neural networks is put forward. This method succeeds in inducing back-propagation neural networks into the ways of computer virus detecting. In comparing with the traditional methods, this new detection is more effective in analyzing system information and file system, and can diagnose which kind of computer virus are infected.

**[Key words]** BP neural networks; Virus detection; Computer virus; Training examples

从1983年计算机病毒首次被确认, 但并没有引起人们的重视。直到1987年计算机病毒才开使受到世界范围内的普遍重视。我国于1989年在计算机界发现病毒。至今, 全世界已发现数万种病毒, 并且还在高速度增加, 其破坏能力也在逐步增强, 已经严重地影响到我们的工作和生活。所以一种确实有效的病毒检测系统对于整个计算机系统是极其重要的。

### 1 当前主要的几种病毒监测方法及存在的问题

#### 1.1 主要的几种检测方法

现今主要的几种流行的检测病毒方法有: 特征代码法、校验和法、行为监测法<sup>[1]</sup>, 这些方法依据的原理不同, 实现时所需开销不同, 检测范围也不同。

##### (1) 特征代码法

特征代码法是使用较为普遍的病毒检测方法。特征码查毒就是检查文件中是否含有病毒数据库中的病毒特征代码<sup>[2]</sup>。这种检测病毒的方法是对于检测已知病毒操作比较简单, 并且开销也比较小。但是采用病毒特征代码法的检测工具必须不断更新版本, 否则检测工具便会老化, 逐渐失去实用价值。病毒特征代码法对从未见过的新病毒就无能为力。

##### (2) 校验和法

将正常文件的内容计算其校验和, 写入文件中保存。定期检查文件的校验和与原来保存的校验和是否一致, 就可以发现文件是否感染病毒, 这种方法叫校验和法, 它既可发现已知病毒又可发现未知病毒。

因为病毒感染并非文件内容改变的唯一的非他性原因, 文件内容的改变有可能是正常程序引起的, 所以校验和法常常误报警。而且此种方法也会影响文件的运行速度。因而用监视文件的校验和来检测病毒, 不是最好的方法。

校验和法也有一些优点, 如方法简单, 能发现一些简单的未知病毒, 也能检测到被查文件的一些细微变化。其缺点是: 对文件内容的变化过于敏感、会误报警、不能识别病毒名称、无法对付隐蔽型病毒。

##### (3) 行为监测法

利用病毒的特有行为特征来监测病毒的方法, 称为行为

监测法。通过对病毒多年的观察、研究, 有一些行为是病毒的共同行为, 而且比较特殊。当程序运行时, 监视其行为, 如果发现了病毒行为, 立即报警。

行为监测法也有一些长处, 如可发现一些简单的未知病毒。行为监测法的短处: 可能误报警、不能识别病毒名称及类型、实现时有一定难度。

#### 1.2 存在的几个问题

当前几种流行的检测手段都无法自动提取病毒特征。整个检测系统无法自动进行检测学习, 没有联想记忆功能, 不能从事大规模并行处理, 也没有实时计算的能力<sup>[1]</sup>。

为了解决这些问题, 提出了一种基于BP神经网络的病毒检测方法。

## 2 基于BP神经网络的病毒检测方法

### 2.1 反向传播神经网络

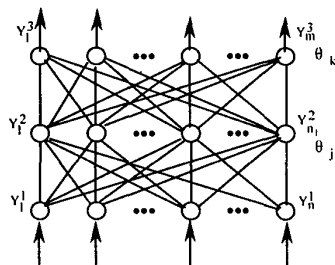


图1 一个三层的BP神经网络拓扑结构

其中:  $y_i^1$  为输入层节点i的输出;  $y_j^2$  为中间层节点j的输出;  $y_k^3$  为输出层节点k的输出;  $T_k$  为输出层节点k对应的教师信号;  $w_{ij}$  为节点i与节点j间的连接权值;  $x_k$  为节点j与节点k间的连接权值;  $\theta_j$  为中间层节点j的阈值;  $\theta_k$  为输出层节点k的阈值。

反向传播神经网络(Back-Propagation Neural Networks, BPNN)又称为BP模型, 其结构图如图1。在这一神经网络模

**基金项目:** 国家自然科学基金资助项目(60064002)

**作者简介:** 郭晨(1979—), 男, 硕士生, 研究方向: 网络安全与网络数据挖掘; 梁家荣, 教授; 梁美莲, 硕士生

**定稿日期:** 2004-01-13 E-mail: bearr@21cn.com

型中引入了中间隐含神经层。故标准的BP模型由3个神经层次组成,其最下层称为输入层,中间层称为隐含层,最上层称为输出层。各层内的神经元之间没有连接<sup>[3]</sup>。

## 2.2 基于BP神经网络的病毒检测系统

### 2.2.1 确定一个输入表示方案

在建立一个病毒检测系统之初，必须对整个系统的实用性进行规划，使之能满足实际工作的需要。主要考虑的问题有：找到一个合适的输入模式方案。

计算机的引导扇区的长度只有512B。因此输入表达方式要想涵括所用的原始位或原始字节是根本不可能的。所以改为使用一种新颖的表达方案,首先利用一些自动程序将计算机病毒切割成3B的碎片,这3B的字符碎片是从一个训练样本集中分析产生的。然后再从这3B的病毒碎片中提取一系列以存在和不存在表示的病毒特征,而这些以存在和不存在表示的病毒特征又通过一系列的“0”和“1”来表示,其中,“1”表示存在,“0”表示不存在。所得的这些病毒特征应该频繁地出现在被感染的病毒扇区中,而不经常出现在一些正常的合法扇区中。

基于BP神经网络的病毒检测系统的输入模式就是：通过以下几种方式得到病毒的输入特征，首先建立一个由3B的字符碎片组成的列表，这个列表要包含所有的在扇区中存在病毒的3B特征字符碎片，其次就是要清除集合中的合法扇区和那些经常出现在独立集合的没被感染的计算机程序的3B病毒特征。接下来就是要从这些3B病毒特征表中分析出每一个病毒的病毒特性(由“1”或“0”表示)。最后形成一个完全的病毒特征表，其中训练集中每种病毒至少有4种特征存在，即其值为“1”。从而就得出我们对BP网络进行训练学习的训练输入样本集。如图2。

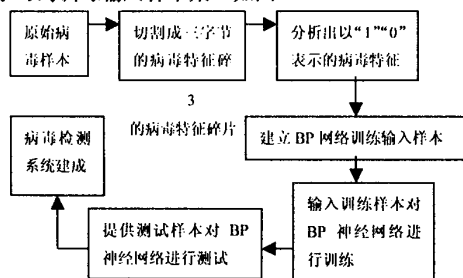


图2 基于BP神经网络的病毒检测系统训练流程图

### 2.2.2 病毒样本训练

在计算机系统中可以用于这种病毒检测工作的训练数据十分有限。在整个系统中所利用的训练数据集只能从约200个被感染的扇区和合法的引导扇区中得出。而在所有的可利用病毒样本中,还必须预留下50%的病毒样本作为以后的测试样本。所以最终的BP神经网络存在一定的“误确定”和“误否定”是肯定存在的,但是系统应该保持在一定的“误确定”和“误否定”范围之内。

另外一个基于训练数据的更深的问题就是：由输入模式结构可知看出，每一个合法的程序或者扇区都应该由一系列“0”组成的输入模式来表示。因此，如果其中的任何一个的分量为“正”，则将会被认为是被病毒感染。显然，这样将会导致诊断异常，那就是当一个合法的程序或者扇区产生了一个错误的正的输入分量时，神经网络将会产生一个误“肯定”。即认为这个合法程序或者扇区是被病毒感染的。

解决这个问题一个可行的办法就是，通过学习训练时认为只包含一个“1”而其他的特征都是“0”的情况下我们仍然认为是合法的程序或者扇区。这样就使得任何一个单一的“1”特征是不能断定是否是被感染的扇区。

### 2.2.3 病毒分类以及病毒测试

在进行样本训练的同时,通过一定的编码规则使得输出结果不仅能检测出病毒的存在,并且能够指出病毒的类型。然后根据病毒的类型提出相应的处理方法。如以下方式:指定输入输出为6维空间,然后建立一种输出一病毒类型映射,如表1所示。

表1 输出特征位-病毒类型映射表

输出特征位	1		2		3		4		5		6	
输出特征值	1	0	1	0	1	1	0	0	1	0	1	0
表示意义	引导型病毒	文件型病毒	MBR型病毒	非MBR型病毒	BR型病毒	非BR型病毒	原码型病毒	非原码型病毒	外壳型病毒	非外壳型病毒	嵌入型病毒	非嵌入型病毒

我们通过在样本训练中引入这个输出特征位-病毒类型映射表,使得BP神经网络在训练中建立起这么一种映射关系,以通过编码的方式达到对病毒分类的功能。然后通过测试样本验证这个输出特征位-病毒类型映射表的正确性。之后就可以在病毒检测中得出病毒的类型,并可以根据病毒的类型给出一定的处理意见。输出特征位-病毒类型映射图见图3。

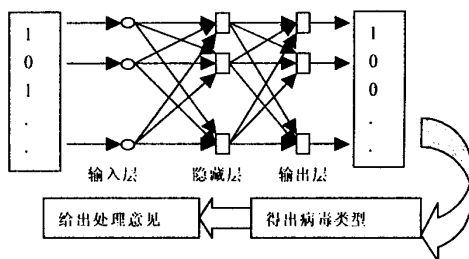


图3 输出特征位-病毒类型映射图

在训练中使用的是BP神经网络，由于病毒样本的不完整，因此用BP神经网络来进行的检测学习只能达到90%~95%的准确率，以此为据，我们可以在隐藏层中设定一定的阈值，以提高检测的准确性。

### 2.3 利用神经网络检测病毒方法的优势

基于BP神经网络启动监测技术是由于通过模仿人的神经细胞来区分被病毒感染和未感染的电脑启动记录,因此该技术提供了更全面的病毒系统检测手段。神经网络接受许多病毒和非病毒的样本,因而学会了辨认病毒,其效果比病毒研究员手工调节的传统启发式技术要好。这个神经网络能自动监测出绝大多数已知、未知的启动记录病毒以及对病毒进行分类处理。

利用BP神经网络实现病毒检测的模式识别时, BPNN能够从输入的数据中, 自动提取病毒特征, 并存储于网络之中。BPNN在处理模式识别问题时, 具有联想记忆, 大规模并行处理, 实时计算和便于硬件实现等特点。

变,这里均以典型的 128b 为例。系统的最大时钟频率为 47MHz,共需 1 228 个可配置逻辑块 (CLB) 和 18 个专用 BLOCK RAM,可同时完成加密和解密功能。加解密之间转换不会使速度产生变化,其吞吐率为 421Mbps。

系统采用内部流水线结构。内部流水线结构相当于把轮运算部件的各功能模块内部插入 D 触发器来划分流水线结构,把功能细化,使每个模块的时延尽可能小,从而提高系统工作频率。这种结构在设计上要求各模块的功能时延(包括运算时延和传输时延)尽可能相等,在采用 CBC、CFB、OFB 反馈加密模式设计时,两组独立的数据必须在加密的任何一个点及时得到。内部流水线结构的速度相对来说要求较高,在资源消耗方面也有一定的限制。系统在加密圈内部增加了流水线寄存器,这使其在基本迭代结构的基础上得到扩充。这样,只要增加时钟频率,就可以同时处理两组数据。这使系统仅在增加部分资源的情况下,加解密的吞吐量几乎是无流水线结构的两倍。

加解密的密钥是并行产生的。圈密钥相对于新一轮的密钥可以在后台产生,可以与前一轮处理数据同步,不会因为密钥的灵活转换而带来速度的降低。能充分利用芯片的面积,利用如资源共享和内部流水线环绕操作来使芯片的面积最小化,使吞吐量和面积之间的比率达到一个最好值。

3 实验结果

系统采用 Model 公司专门为各逻辑器件制造厂商设计的第三方专用仿真工具 ModelSim 进行仿真,它对 VHDL、Verilog 语言的硬件设计仿真方面非常出色。Xilinx 公司的联合仿真环境 ISE 为它提供了无缝接口。本文给出了最后的时序仿真图。

(1) 密钥加载

图 3 是密钥加载过程的时序图,图中所示为 128b 的密钥在 8 个时钟周期中传输,每次传输率为 16b。在一个最小时钟周期内密钥加载是这样生成的: InputValid 信号输入产生了一个高电平(LoadKeyBusy)和一个低电平(AES-Ready),表明在输入有效时在外部显示密钥交换和扩展正在进行。在 8 个有效传输完成之后, LoadKeyBusy 信号产生低电平,表明密钥加载完成。然后密钥扩展开始,产生内部圈密钥。最

(上接第 153 页)

3 结束语

基于神经网络的病毒检测系统技术如今已经在国内外的一些大型的杀毒防毒软件中得到应用。如 1998 年 Symantec 公司宣布把 IBM 公司专利的神经网络引导检测技术集成到诺顿防病毒(Norton AntiVirus)产品中。这种神经网络技术利用人工智能检测引导型病毒,将使 Symantec 公司革命性的 Bloodhound 启发式探索技术更加完善。Norton AntiVirus 中的这种启发式检测技术,可以检测出高达 90% 的新型和未知的引导型病毒。

但是这种基于神经网络的病毒检测系统技术,目前所用的还仅仅局限于一般的单层神经网络,并且缺乏对病毒类型

后整个过程完成之后 AES-Ready 信号变成高电平。

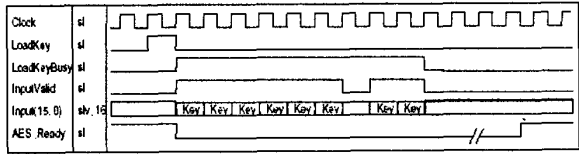


图 3 密钥加载时序图

(2) 数据输入和输出

图 4 是数据输入输出过程的时序图,图中所示为 128b 的密钥在 8 个时钟周期中传输,每次传输率为 16b。在密钥扩展完成之后, AES-Ready 变成了高电平。AES-Start 变为高电平准备加解密,在 AES-Ready 完成 8 个时钟之后, InputValid 信号产生一个低电平,表示一个周期的数据传输完成。在 AES-Ready 经过一个时钟单元的跳变后,重新置为高电平开始新的数据传输。

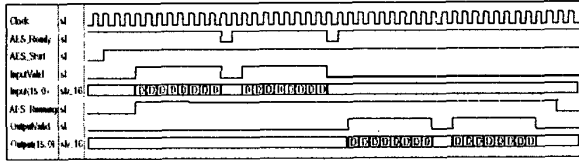


图 4 数据传输时序图

4 结束语

整个设计结构简单、适用性强、加密速度快,它不仅适用于 AES 算法,经改进也可以被用于 3 重 DES 和其它如 Mars 和 RC6 等一些算法。本 IP 核也可以被用作硬件实现中的一部分或者许多主要安全协议的混合应用中,如 IPSEC、SSL、IEEE、802.11A,以及 ATM 的安全应用。

参考文献

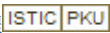
1 AES Home Page[EB/OL].http://www.nist.gov/encryption/aes,2002  
2 Daemen J, Rijmen V. AES Proposal: Rijndael[M]. 1999-09  
3 XLINX Data Book 2000. The Programmable Logic[P]. 2000  
4 侯伯亨, 顾 新. VHDL 硬件描述语言与数字逻辑电路设计. 西安: 西安电子科技大学出版社, 1997

的检测功能,也没有利用到 BP 网络的联想记忆功能,并没有充分发挥出神经网络的检测优势,所以基于 BP 神经网络的病毒检测系统在病毒检测系统中将有广泛的应用前景。

参考文献

1 White R. Open Problem in Computer Virus Research. Virus Bulletin Conference, 1998-10-22  
2 袁忠良. 计算机病毒防治实用技术. 北京: 清华大学出版社,1999  
3 李学桥,马 莉. 神经网络. 工程应用.重庆:重庆大学出版社,1996  
4 李 洗. 计算机病毒与 DOS 奥秘. 贵阳: 贵州科技出版社,1992

# 基于BP神经网络的病毒检测方法

作者: 郭晨, 梁家荣, 梁美莲  
作者单位: 广西大学计算机与信息工程学院, 南宁, 530004  
刊名: 计算机工程   
英文刊名: COMPUTER ENGINEERING  
年, 卷(期): 2005, 31(2)  
被引用次数: 5次

## 参考文献(4条)

1. White R Open Problem in Computer Virus Research 1998
2. 袁忠良 计算机病毒防治实用技术 1999
3. 李学桥; 马莉 神经网络. 工程应用 1996
4. 李洗 计算机病毒与DOS奥秘 1992

## 本文读者也读过(5条)

1. 梁玲 宏病毒感染过程及防范措施研究[期刊论文]-科技情报开发与经济2009, 19(30)
2. 王志斌, 陈文梅, 褚良银, 王升贵, WANG Zhi-bin, CHEN Wen-mei, CHU Liang-yin, WANG Sheng-gui 基于MATLAB的BP神经网络在旋流器模拟设计中的应用[期刊论文]-流体机械2007, 35(10)
3. 危胜军, 胡昌振, 姜飞, WEI Shengjun, HU Changzhen, JIANG Fei 基于BP神经网络改进算法的入侵检测方法[期刊论文]-计算机工程2005, 31(13)
4. 唐志芳, 时海涛, 鲁华祥, 王守觉, TANG Zhifang, SHI Haitao, LU Huaxiang, WANG Shoujue 基于BP神经网络的系统级电源管理算法[期刊论文]-计算机工程2006, 32(4)
5. 陈丽红, 甘祥根, 杨冬保 基于Matlab神经网络连阴雨预报模型研究[会议论文]-2009

## 引证文献(5条)

1. 李天志, 王海涛, 徐凤生 基于进程管理的计算机病毒检测系统[期刊论文]-福建电脑 2006(10)
2. 叶清, 吴晓平, 程晋 基于规则优化与排序的恶意代码匹配检测[期刊论文]-海军工程大学学报 2010(4)
3. 李静, 周跃进 复杂网络病毒防治系统设计与实现[期刊论文]-计算机应用与软件 2008(2)
4. 郝向东, 王开云 典型恶意代码及其检测技术研究[期刊论文]-计算机工程与设计 2007(19)
5. 谢金晶, 张艺濒 基于改进的K-最近邻算法的病毒检测方法[期刊论文]-现代电子技术 2007(3)

本文链接: [http://d.wanfangdata.com.cn/Periodical\\_jsjgc200502057.aspx](http://d.wanfangdata.com.cn/Periodical_jsjgc200502057.aspx)