

基于心跳行为分析的木马快速检测方法

孟磊, 刘胜利, 刘龙, 陈嘉勇, 孙海涛
(解放军信息工程大学信息工程学院, 郑州 450002)

摘要: 基于通信行为分析的木马检测算法的计算复杂度较高。为此, 提出一种基于心跳行为分析的木马快速检测方法, 通过对木马通信中心跳行为的描述, 选取 2 个会话特征对木马通信流与正常通信流进行分类, 基于该方法设计一个木马快速检测系统 TRDS。实验结果表明, TRDS 能够在百兆线速网络中快速有效地检测出木马通信。
关键词: 木马检测; 会话特征; 通信流分析; 行为分析; 心跳行为; 快速检测

Trojan Rapid Detection Method
Based on Heartbeat Behavior Analysis

MENG Lei, LIU Sheng-li, LIU Long, CHEN Jia-yong, SUN Hai-tao
(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002, China)

【Abstract】 Trojan detection algorithm based on behavior analysis of communication has high computational complexity. Addressing the problem, this paper proposes a Trojan rapid detection based on heartbeat behavior analysis. The method selects two session attributes to describe the difference between Trojan communication flow and normal communication flow on the basis of description of heartbeat behavior in the Trojan communication large numbers of analysis on Trojan samples. And then Trojan Rapid Detection System(TRDS) is built based on the method. Experimental results show that TRDS can detect the Trojan communication in the 100 Mbit/s network rapidly and efficiently.
【Key words】 Trojan detection; session feature; communication flow analysis; behavior analysis; heartbeat behavior; rapid detection
DOI: 10.3969/j.issn.1000-3428.2012.14.004

1 概述

随着网络的高速发展, 网络安全也越来越受人关注。木马技术作为黑客应用的关键技术之一, 对网络安全构成了严重威胁。增强对木马通信的检测, 对于提升网络安全有着重要的意义。目前, 木马通信检测技术主要采用特征码匹配技术^[1]。相较于基于特征码匹配的检测技术, 基于通信行为分析的检测技术在时效性和扩展性方面具有明显优势, 有利于发现潜在的、未知的网络窃密行为和威胁, 具有更广的应用前景^[2-3]。

然而基于通信行为的木马检测算法通常存在计算复杂度较高的问题, 在实时监控应用中给监控系统带来庞大的计算开销^[4]。因此, 如何设计具有高检测性能且计算复杂度较低的检测算法, 进而实时有效地检测木马的网络通信行为就成为当前一个重要的理论和技术问题。由此, 本文通过分析木马通信中的心跳行为, 提出一种木马快速检测方法。

2 心跳行为描述

2.1 心跳行为

目前, 大多数木马采用数据流伪装方式隐藏通信行为来躲避网络安全设备的检测, 如采用 HTTP 隧道技术^[5]等。目前对于此类隧道技术的安全检测效果较为薄弱。但由于木马通信本身的特殊性, 仍然存在一些行为特征不能做到和正常通信行为保持完全一致, 通过分析其差异仍然可以区分出木马通信。

本文通过对大量木马样本进行分析, 发现木马通信中存在一些包含特定信息的数据包来表明被控主机的网络存活性。这类数据包之间虽然存在着一定的差异, 但木马设计者考虑到网络环境差异、被控主机系统不同和可操作性等问题,

通常会采用通用的解决方法, 由被控端向控制端发送包含特定信息的数据包, 这些信息通常包括被控主机的主机名、IP 地址、操作系统类型及语言版本和攻击者设定的标识等。由于这些包含特定信息的数据包的目的是表明被控主机的网络存活性, 且这些数据包的发送时间一般具有一定规律, 类似于人类的心脏跳动, 因此称之为心跳包。

在木马实现过程中, 通常心跳包并不是单独存在的, 还伴随一些表示确认的数据包, 整个过程称为心跳过程, 相邻 2 次心跳过程间的时间间隔称为心跳间隙。由连续心跳过程和心跳间隙交替出现组成的通信行为称为心跳行为。

如图 1 所示, 是一款木马样本心跳行为的数据包。方框标示的一组数据包为一次木马心跳的心跳过程。

No.	Time	Source	Destination	Protocol Info
25	72.781767	192.168.1.144	192.168.1.145	SSL Continuation Data
26	72.913169	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=236 win=64005 Len=0
27	72.917877	192.168.1.144	192.168.1.145	SSL Continuation Data
28	73.131805	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=248 win=63993 Len=0
29	132.786691	192.168.1.144	192.168.1.145	SSL Continuation Data
30	132.958311	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=252 win=63989 Len=0
31	132.958414	192.168.1.144	192.168.1.145	SSL Continuation Data
32	133.177305	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=264 win=63977 Len=0
33	192.784392	192.168.1.144	192.168.1.145	SSL Continuation Data
34	192.958032	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=268 win=63973 Len=0
35	192.988185	192.168.1.144	192.168.1.145	SSL Continuation Data
36	193.116679	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=280 win=63961 Len=0
37	252.787398	192.168.1.144	192.168.1.145	SSL Continuation Data
38	252.943226	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=284 win=63957 Len=0
39	252.943408	192.168.1.144	192.168.1.145	SSL Continuation Data
40	253.182216	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=296 win=63945 Len=0
41	312.787370	192.168.1.144	192.168.1.145	SSL Continuation Data
42	312.963223	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=300 win=63941 Len=0
43	312.993226	192.168.1.144	192.168.1.145	SSL Continuation Data
44	313.211422	192.168.1.145	192.168.1.144	TCP https > wfrmotertm [ACK] Seq=5 Ack=312 win=63929 Len=0

图 1 木马心跳行为的数据包

基金项目: 郑州市科技创新团队基金资助项目(10CXTD150)
作者简介: 孟磊(1987—), 男, 硕士研究生, 主研方向: 网络安全; 刘胜利, 副教授; 刘龙, 助教; 陈嘉勇, 博士研究生; 孙海涛, 助理工程师
收稿日期: 2011-10-19 **E-mail:** menglei1314@gmail.com

整个心跳过程通信共包含 4 个数据包, 通过时间间隔大小可以明显发现存在于心跳过程之间的心跳间隙。

心跳行为是由若干心跳过程和心跳间隙交替出现组成的, 如图 2 所示。心跳过程一般由 3 个以上的数据包组成, 特殊情况的也具有 2 个数据包会话, 双方一边一个。本文选择的心跳行为的检测重点在于检测有规律的心跳间隙存在, 所以, 基于心跳行为的木马通信检测在实际应用中可以支持单向数据流检测以及抵抗部分数据丢包导致的漏报。

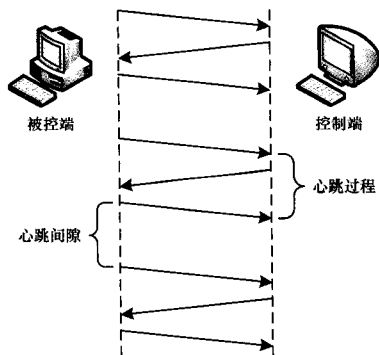


图 2 保持连接阶段

2.2 特征选取

本文通过对 80 个木马样本的分析, 发现木马通信与正常网络通信相比有以下 2 个明显的会话统计特征:

(1) 心跳过程接收和发送的数据包比例。

不同木马在心跳形式上存在着部分差别, 但是对于同一款样本而言, 其心跳过程是完全一致的。因此, 木马在每个心跳过程中行为是一致的, 发送和接收的数据包的比例也是恒定的。当部署在单向数据流上时, 心跳过程表现为恒定的数据包数。当部署环境的数据不够稳定时, 可以通过调节心跳过程的相似度判断阈值来降低漏报。在降低心跳过程相似度判断时可能会导致部分虚警率提高, 但目前实际环境测试中尚未发现此类虚警干扰。

(2) 心跳间隙的时间长度。

不同木马的心跳间隙的时间长度存在较大差别, 但对于同一款样本而言, 大部分样本的时间长度为定值。然而木马的心跳间隙并不像心跳过程一样恒定不变, 为了躲避统计分析, 部分木马采用了可变的心跳间隙, 其目的是用变化的心跳间隙来隐藏恒定的心跳过程, 使其变得无规律可循。在正常的网络通信过程中, 访问发起受到用户的控制, 发起时机完全是随机的。但是木马通信以保证其稳定性为首要任务, 因此, 所使用各种算法通常也只是产生伪随机的心跳间隙, 与正常网络通信仍然存在一定的差别。

2.3 特征分析

本文利用时频分析来验证选取检测特征的分类效果。原始数据包时间间隔采样结果记为: $X = \{x_i\}_{i=1}^n$, 其中, X 表示数据包时间间隔采样集合; x_i 表示具体的采样值。离散傅里叶变换 (Discrete Fourier Transform, DFT) 变换公式为: $DFT(X) = \{y_i\}_{i=1}^n$, 其中, y_i 表示经过 DFT 变换后的采样值。经过 DFT 变换后的心跳间隙和正常网络通信的数据包时间间隔如图 3~图 8 所示。图 3~图 6 表示木马心跳间隙采样数据的 DFT 变换结果, 其中, 图 3、图 4 表示采用定时长方式进行心跳的木马样本, 而图 5、图 6 表示采用变时长方式进行心跳的木马样本。图 7、图 8 表示正常网络通信数据包时间间隔的 DFT 变换结果。

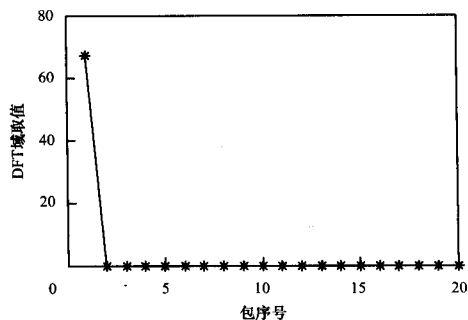


图 3 Bitfrost 木马 DFT 采样图

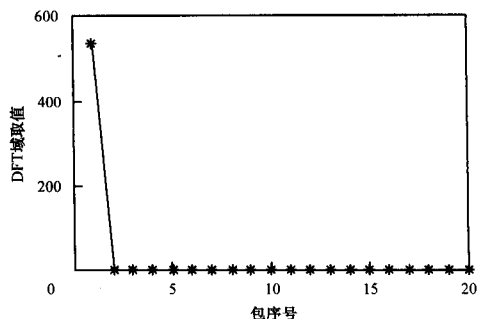


图 4 Gh0st 木马 DFT 采样图

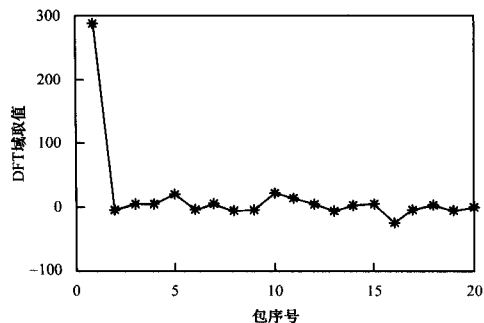


图 5 变长心跳木马 1DFT 采样图

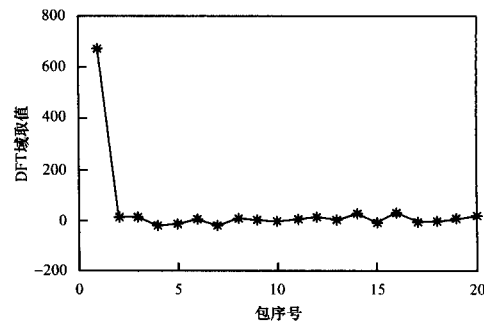


图 6 变长心跳木马 2DFT 采样图

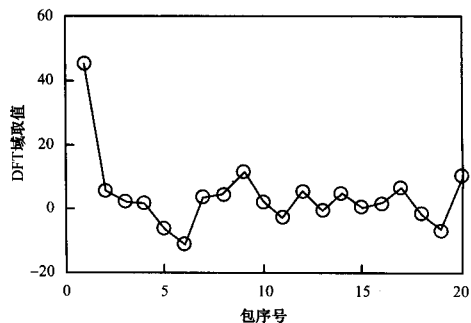


图 7 正常网络回话 1DFT 采样图

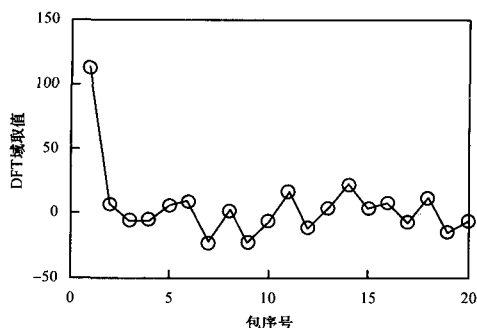


图8 正常网络会话 2DFT 采样图

分析结果表明: 正常网络通信数据包的时间间隔转换到频域后, 中频和高频系数都比较大, 这表明正常网络通信数据包的时间间隔表现出非平稳信号的特性, 这与正常网络通信的随机性相符。木马心跳间隙则相反, 由于木马心跳具有一定的规律, 导致其表现出相对平稳信号的特性。其中, 采用定时长方式进行心跳的木马, 由于心跳规律非常明显, 使得其信号的中高频系数几乎为 0, 而变时长心跳的木马, 采用了各种伪装的方式, 其信号特性虽表现出一定的波动性, 但与正常网络通信相比还是较为平稳。因此, 可以通过判断心跳间隙的中频和高频系数的值域来实现木马检测, 从而验证了前文选取行为特征在木马检测分类中的有效性。

3 检测实现

本文检测首先通过查询网络会话列表, 将捕获的网络数据按会话进行分类, 然后对会话数据进行频域计算, 得到心跳间隙的中频和高频系数, 根据阈值判断是否为木马会话。由于网络会话数量大, 在千兆网络流量环境下按照常规查询方式会出现数据丢包, 影响检测效率。同时频域计算复杂度较高, 导致对计算资源占用过多而影响系统性能。为了解决上述 2 个问题, 本文采用如下 2 种方法。

3.1 查询优化

本文选取四元组(源 IP, 目的 IP, 源端口, 目的端口)作为网络会话唯一标识。每个数据包通过四元组查找到相应的会话, 将信息添加到网络会话链表中。

一般来说, 四元组可以使用多维数组或多级链表进行保存。数组结构具有存储效率高、查找方便、存取速度快等优点, 但是数组要求预先为其分配存储空间, 一旦建立则无法改变数组大小, 容易造成空间浪费。但是网络会话数量不固定, 因此无法为其预先分配空间。链表的优点是可动态添加或删除、不需要预先分配空间, 但缺点是查找速度慢。因此, 本文考虑采用数组链表——数组和链表相结合的数据结构来记录会话的四元组, 牺牲一定的存储空间来提高查找效率。本文参照分块查找算法思想采用四级链表的数据结构来查询网络会话的四元组, 然后参考哈希表查找算法思想将哈希表作为多级链表的索引来提高查找效率。

本文发现如下规律: 对分布均匀的分量使用数组链表结构, 对分布不均匀的分量使用链表结构, 这样可以获得更高的查找效率。

为便于证明, 本文以数组链表结构为例进行分析如下:

设会话数量为 S , 若将所有的会话以单链表的形式进行组建, 每次系统接收到数据包后都要对会话链表进行顺序查找(顺序查找的平均时间复杂度为 $O(\frac{S}{2})$)。

以数组链表的形式记录会话, 设数组长度为 n , 数组的

第 i 个节点下的会话链表个数为 α_i , 则产生第 i 个节点的概率为 $\frac{\alpha_i}{S}$ 。

对链表进行查询的平均时间复杂度为:

$$O\left(\frac{\alpha_i}{S} \cdot \frac{\alpha_i}{2}\right) = \frac{O(\alpha_i^2)}{2S}$$

根据定理均方根大于等于算术平均数可得:

$$\sqrt{\frac{\sum_{i=1}^n \alpha_i^2}{n}} \geq \frac{\sum_{i=1}^n \alpha_i}{n} = \frac{S}{n}$$

将不等式两边同时平方可得:

$$\sum_{i=1}^n \alpha_i^2 \geq \frac{S^2}{n}$$

当且仅当 $\alpha_1 = \alpha_2 = \dots = \alpha_n$ 时, 其中, $\sum_{i=1}^n \alpha_i = S$ 。即 $\alpha_i = \frac{S}{n}$

时 $\sum_{i=1}^n \alpha_i^2$ 最小。

由此可知, 当数组等分链表时查找的时间复杂度最低:

$$O\left(\frac{S}{2n}\right), \text{ 该时间复杂度小于单链表的查找时间复杂度。}$$

在具体实现时, 为获得更好的查找效率, 本文对分布均匀的分量使用哈希表建立索引以提高查询效率, 对分布不均匀的分量使用链表结构以节省空间占用。

以下为四元组中各元素在会话标示中的含义及表示范围:

(1)源 IP 地址。源 IP 地址为待保护内网主机 IP 地址。相对互联网而言, 内网 IP 地址空间小, 且分布均匀。

(2)源端口。根据协议规定, 源端口一般使用 1024~65535 之间的任意端口。

(3)目的 IP 地址。目的 IP 地址的取值范围为整个 IPv4 地址空间, 范围巨大, 而且分布无序。

(4)目的端口。目的端口一般为协议的制定端口, 范围主要集中在 1~1023 之间, 大部分网络通信的目的端口为 80、443、8080 等协议端口。

根据上述分析, 按源 IP 地址、源端口、目的 IP 地址、目的端口的顺序分别建立 4 级链表。其中, 源 IP 地址取值范围较小, 而且分布均匀, 尤其最后 1 Byte 的分布更为均匀, 而且一般都是连续分配, 其分布范围为 1~254。因此, 本文最终选取源 IP 地址的最后 1 Byte 进行哈希值的计算, 建立哈希表。在理想状态下, 通过此哈希表优化后, 查找时间复杂度为原来的 $1/254$, 系统效率得到明显提升。用于提取会话的链表结构如图 9 所示。

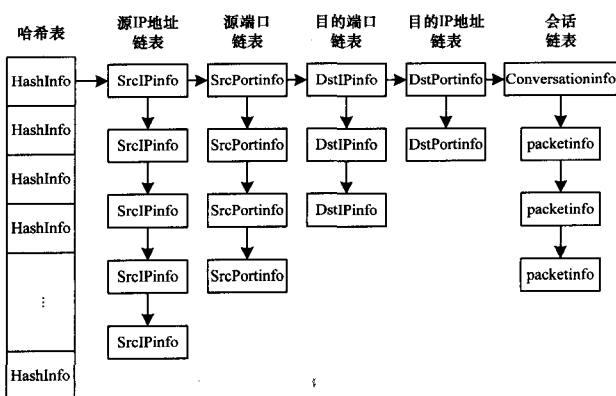


图9 网络会话表结构

3.2 检测优化

本文检测方法主要依靠判断心跳间隙的中频和高频系数取值范围来实现。从而先定义心跳间隙的平稳性如下：

$$Stability = \frac{\sum_{i=2}^n |y_i|}{n-1} \leq \omega$$

当平稳性小于阈值 ω 时，则判定其为木马通信。

由于频域系数的计算复杂度较高，本文基于二次 haar 小波分解的本质构造了计算复杂度较低的可检测变长心跳包的统计量。

$$\text{记 } w_{2i} = \frac{x_{2i} - x_{2i-1}}{2}, \quad w_{2i}' = \frac{w_{2i} - w_{2i-1}}{2}; \text{ 则有:}$$

$$w_{4i}' = \frac{w_{4i} - w_{4i-2}}{2} = \frac{x_{4i} - x_{4i-1} - x_{4i-2} + x_{4i-3}}{4}$$

其中， w_{4i}' 相当于二次 haar 小波分解后的高频系数。此时平稳性定义为：

$$Stability = \frac{x_{4i} - x_{4i-1} - x_{4i-2} + x_{4i-3}}{4} \leq \omega$$

当平稳性小于阈值 ω 时，则判定其为木马通信。

4 计算复杂度分析

本文检测方法主要分为特征提取和检测 2 个阶段，下文分别进行计算复杂度分析。

(1) 假设会话包含 n 个数据包，对特征提取阶段进行计算复杂度分析如下：

提取心跳接收数据包数量和发送数据包数量之比的计算复杂度为 $O(n)$ 。由于提取心跳间隙平稳度时采用 2 次 haar 小波分析，该算法的计算复杂度为 $O(4n)$ ，因此综合考虑得到提取心跳间隙平稳度的计算复杂度为 $O(4n)$ 。

(2) 对检测阶段进行计算复杂度分析如下：

在连接保持无操作阶段，对心跳间隙平稳度进行判断的计算复杂度为 $O(1)$ 。将心跳期接收数据包数量和发送数据包数量的比值保存在数组结构中，假设数组结构长度为 n ，则判断该特征属性的计算复杂度为 $O(n-1)$ 。

综上所述，本文所选取的行为特征的最差计算复杂度为 $O(4n)$ ，因此，本文方法效率较高。

5 检测原型

为了验证本文木马通信检测方法的高效性，设计实现了木马快速检测系统(Trojans Rapid Detection System, TRDS)。该系统主要分为数据采集模块、特征提取模块、木马判定模块和报警响应模块，其结构如图 10 所示。

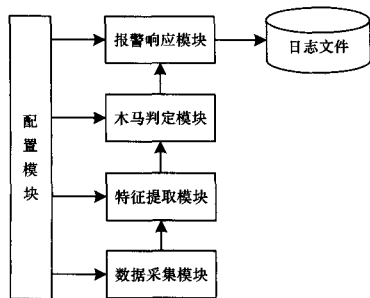


图 10 TRDS 结构

在该系统实现中，数据采集模块采用 winpcap 来实现数据捕获。特征提取模块负责查询会话链表，将每个数据包按会话分类计算特征值。木马判定模块遍历会话链表，判断会话特征是否满足木马特征。对于符合心跳特征的会话，由报警响应模块提取相应信息记入日志文件。配置模块用于对其

他各个模块进行配置。

6 实验分析

为了获得被保护网络中完整的网络数据，进行检测分析，本文实验中通过将运行 TRDS 的 PC 机(Intel 酷睿 i5 760 四核 2.8 GHz/4 GB Memory/1 TB HardDisk, 安装 Windows 2003 Server 操作系统)部署在内网和互联网之间的交换机镜像端口，获得检测数据。环境部署如图 11 所示。

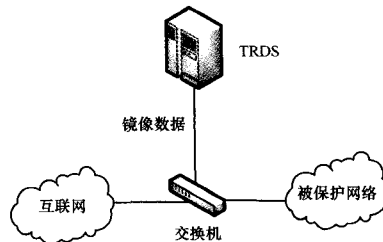


图 11 TRDS 部署

实验数据来源于由 60 台上网主机组成的百兆局域网。在其中 1 台主机上进行木马植入，测试木马样本 107 个(其中包括分析时的 80 个样本以及 27 个未分析样本)，控制端设立在局域网外。

表 1 为 TRDS 检测到的存在心跳行为的 101 个木马样本。

表 1 检测结果

木马样本数	检测出样本数	未检出样本数	检测率/(%)
107	101	6	94.4

表 2 列出测试样本中常见的 11 种木马样本心跳特征。

表 2 部分木马样本的心跳特征

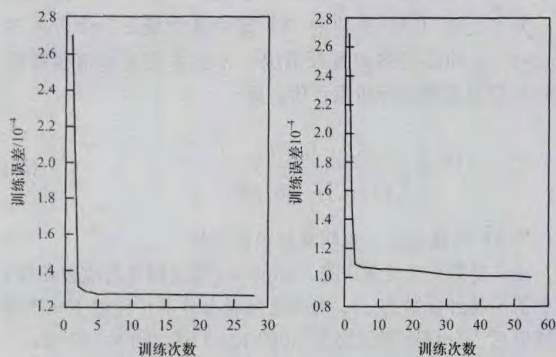
木马名称	心跳时间间隔/s
灰鸽子	30.1
Gh0st	120.0
Pcshare	34.2
流萤	无
Cmder	30.0
任我行 netsys	30.0
Biforest	34.1
KPR 远程控制器	34.1
炽天使远程控制系统	34.1
黑客防线新一代	60.2
木马帝国远程控制软件	34.1

实验结果表明，本文提出方法在百兆线速环境中可以较好地检测出木马通信，并且对于未被分析的木马样本同样可以检测。在漏检情况中，以流萤木马最为代表性，是由于没有稳定的保持连接而产生漏报。但这类木马也因为没有稳定的保持连接导致木马控制不稳定，现在已经很少使用。

7 结束语

本文通过分析木马通信行为的特点，选取木马心跳行为用于木马检测。该行为定位于木马建立连接后的稳定通信阶段，近年来应用于木马的新技术，如 P2P 技术、Fast-Flux 技术等大部分应用于反弹建立连接阶段，对本文采用的检测方法并未造成影响。该方法通过会话特征统计，实现对正常通信流和木马通信流进行分类。在检测能力上，不仅能对现有的木马进行有效检测，还能检测网络中出现的具有类似通信特征的新型木马。在方法实现上，本文方法的特征提取简单且算法实现复杂度低，方便在大流量网络中检测使用。

(下转第 20 页)



(a)第1类 Chebyshev 神经网络最佳拓扑结构下的收敛曲线 (b)第2类 Chebyshev 神经网络最佳拓扑结构下的收敛曲线

图4 Chebyshev 神经网络最佳拓扑结构下的收敛曲线



图5 不同算法的 lena 复原图

为了更好地说明 Chebyshev 神经网络的性能及复原效果,将各种复原效果图的性能列于表1。从表中可以看出,基于 Chebyshev 神经网络及其衍生算法复原图的 PSNR 值大

于 L-M 优化的 BP 神经网络及文献[2-3]中的算法。

表1 不同算法的 PSNR 比较

图像	PSNR	图像	PSNR
图 5(b)	17.894	图 5(f)	27.362
图 5(c)	27.131	图 5(g)	27.452
图 5(d)	26.769	图 5(h)	27.419
图 5(e)	25.641	图 5(i)	27.661

5 结束语

Chebyshev 神经网络隐层神经元的激励函数为一组 Chebyshev 正交三角基函数,能够自适应调整模型结构,且具有良好的非线性逼近能力。第1类 Chebyshev 神经网络只需调节隐层至输出层一个环节的值,加快了收敛性。而第2类 Chebyshev 神经网络能较好地适应动态系统要求。本文通过实验得出了2类 Chebyshev 神经网络及相应的衍生算法的复原图像,与其他算法相比,具有较好的效果。

参考文献

- [1] 周玉,彭召意.运动模糊图像的维纳滤波复原研究[J].计算机工程与应用,2009,45(19):181-183.
- [2] Wu Yadong, Zhu Qingxin, Sun Shixin. Variational PDE Based Image Restoration Using Neural Network[J]. Image Processing, 2007, 1(1): 85-93.
- [3] Gan Xiangchao, Liew A W, Yan Hong. A POCS-based Constrained Total Least Squares Algorithm for Image Restoration[J]. Journal of Visual Communication and Image Representation, 2006, 17(5): 986-1003.
- [4] 邹阿金,张雨浓.基函数神经网络及应用[M].广州:中山大学出版社,2009.
- [5] 张雨浓,陈裕隆,姜孝华,等.陈一种权值直接确定及结构自适应的 Chebyshev 基函数神经网络[J].计算机科学,2009,36(6): 210-213.
- [6] 丛爽,向微. BP 网络结构、参数及训练方法的设计与选择[J]. 计算机工程,2001,27(10): 36-38.
- [7] Kar P S I, Jha A. On-line System Identification of Complex Systems Using Chebyshev Neural Networks[J]. Applied Soft Computing, 2007, 7(1): 364-372.
- [8] Mishra S K, Panda G, Meher S. Chebyshev Functional Link Artificial Neural Networks for Denoising of Image Corrupted by Salt and Pepper Noise[J]. International Journal of Recent Trends in Engineering, 2009, 1(1): 413-417.

编辑 任吉慧

(上接第16页)

参考文献

- [1] Zhang Like, White G B. An Approach to Detect Executable Content for Anomaly Based Network Intrusion Detection[C]//Proc. of Parallel and Distributed Processing Symposium. Long Beach, USA: [s. n.], 2007: 1-8.
- [2] Dusi M. Tunnel Hunter: Detecting Application-layer Tunnels with Statistical Fingerprinting[J]. Computer Networks, 2009, 53(1): 81-97.
- [3] Liu Ting, Guan Xiaohong, Zheng Qinghua, et al. Prototype Demonstration: Trojan Detection and Defense System[C]//Proc. of the 6th IEEE Conference on Consumer Communications and Networking Conference. Piscataway, USA: [s. n.], 2009: 64-65.
- [4] 郭修昌. 基于决策树的网络隐蔽通道检测模型的研究[D]. 南京: 南京理工大学, 2009.
- [5] 孙海涛, 刘胜利, 陈嘉勇. 基于操作行为的隧道木马检测方法[J]. 计算机工程, 2011, 37(20): 123-126.

编辑 任吉慧