基于文件静态特性评估的特洛伊木马检测机制

The Detection of Trojan Horse Based on the File's Static Characteristics

王维 1 肖新光 2 李柏松 3 戴敏 4

1:安天实验室（david@antiy.net）2:安天实验室（seak@antiy.net）

3:安天实验室（swordlea@antiy.net）4:天津理工大学（daimin@eyou.com）

摘要

木马等后门程序对计算机用户的威胁已众所周知，对于未知木马的查杀能力现已成为各大安全厂商致力研发的热点。本文提出一种新的检测机制，针对木马的静态信息进行分析，建立了基于 BP 神经网络的数据处理模型。

PE 文件是 Windows 平台下的可执行文件，也是木马等后门程序的主要载体，由于 PE 文件的结构复杂、灵活，而 BP 神经网络模型的自身要求，导致数据的建模的过程遇到两个难点，我们首先引入了变化粒度的概念，解决了输入端非线性数据的问题。关于如何将 PE 文件中大量的字符串归 1 化(（0，1）区间)，我们这里使用了统计学的理论，巧妙地绕过数据类型的直接转化，而不丢失 PE 文件本身的的特征。最终我们建立了一套较为成功的解决方案，取得了较为理想的效果。

关键字：木马，神经网络，PE 文件，网络安全

Abstract

It is well known for us to know the threaten of Trojan like programs causes, and it is popular for the security vendor to work for the detection of the Trojan program . Here we offer a new detection method of Trojan, and it is based on the ANN(Artificial Neural Network) theory through the analysis of the Trojan program’s file itself.

PE file is short for Portable Executable File on the Windows platform, and it is the main carrier of Trojan program. We have encountered two problems while building the data model, since the PE file’s format is quite complex and facility, and for the ANN itself ’s specialties request. First, we solve the problem of nonlinear input of PE files through bring the idea of changing granularity of the data; And then, we unified the string that is occurred a lot in PE file by using the statistics theory without losing any information. We finally made a successful blue print, and reached a satisfied result.

Key Words：Trojan horse, Artificial Neural Network,PE(Portable Executable) File,Network Security

一、 静态评估方法与传统检测机制

针对以往传统的特征码检测技术，有其局限性，我们事先必须获得大量的样本，从中提取出正确的特征码，最终建立一套较为完备的样本库。事实证明，这个环节是必需的。而用户必须通过某种途径经常更新其本地病毒库。然而这种传统检测机制的准确性是毋庸置疑的。

我们这里提到的未知木马的检测技术其准确性往往达不到要求，但由于其可对已知样本数据进行分析，学习，从而提取出区别于绝大部分正常文件的特征属性， 最终达到预测的目的。

现在我们将传统检测与静态评估比较如下：

	传统检测	评估检测
病毒库文件	需要	不需要
程序更新	基本不需要	需要
经常更新下载	需要	不需要
更新下载大小	几百 K	<1K
检测未知样本	基于特定短码搜索	目标所在
准确性	100%	？？%
查杀速度	快	？？

由于此项目的研发工作尚处于实验室阶段，保守起见，我们没有给出准确性和查杀速度的实际值。

二、 问题的归类与算法的选择

我们的要求是，从众多文件中检测出木马，如果我们将 Windows 下 PE 文件分为木马文件和正常文件，那么从最终目的上来看，这实际上是辨别一个 PE 文件是木马文件还是正常文件的分类问题，关于分类问题的算法有很多，因为考虑到样本的信息量非常大，并且直观上信息从本身并不能看出其相关性，这就要借助某种现代算法进行学习训练，从而达到预测未知信息的目的。而 BP 神经网络是解决这个问题的代表，并且已经有了成熟的理论基础，以及前人总结的丰富经验，所以我们选择了神经网络作为设计的主要算法，通过感知机的原理学习 PE 文件的静态特性，从而达到检测未知木马的目的。

三、 样本选择与采集

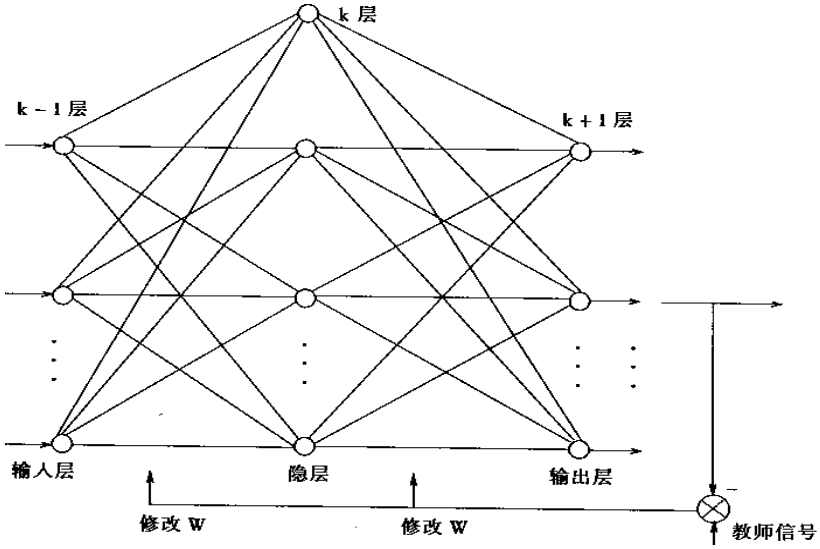
样本采集的好坏直接影响到整个实验的结果，所采集的样本具有一定的普遍性和代表性，理论值与实际值差别则不会有很大出入；如果样本带有局部性，脱离了一般性，即使得出结论，也会导致实验结果与实际产生很大的偏离。本课题的根本目的是要找出正常文件与木马文件的差异，也就是木马文件的静态特性，所以在选择样本时，既要选择有代表性的 PE 文件，又要满足一定的数量。而且在 Windows 下，这些文件要具有普遍性，所以我们选择正常文件的方法是在 Windows 搜索所有的 Exe， DLL 等 PE 文件，然后从中随机挑出 228 个文件，其中 154 个 Exe 文件，74 个 DLL 文件。我们的木马样本是从中国安天实验室样本库中随机提取了 158 个木马样本，然后依据这些数据进行了后续的实验。

四、 误差反传训练算法（BP 算法）

前文提及的学习机制可以直接用来对单层网络进行训练，如感知器学习。在多层网络中，因网络出现了隐含层，隐含层的输出并没有教师信号可供修正权系数。1986 年 Rumelhart 提出了反向传播学习算法，即 BP(back propagation)算法。这种算法可以对网络中各层的权系数进行修正，故适用于多层网络的学习。由于这种算法在本质上是一种神经网络学习的数学模型，所以有时也称为 BP 模型（本文没有严格区分算法和模型两者的异同）。BP 算法是为了解决多层前馈神经网络的权系数优化而提出来的，所以 BP 算法也通常暗示着神经网络的拓扑结构是一种无反馈的多层前馈网络，因此，有时也称多层前馈网络为 BP 网络。

BP 算法的原理简述

BP 算法是用于多层前馈网络的学习算法，多层前馈网络的结构如下图所示：



它包括输人层、输出层以及处于输入输出层之间的中间层。中间层可以为单层或多层，由于它们和外界没有直接联系，故也称为隐含层，隐含层中的神经元称隐单元。隐含层虽然和外界不连接，但它们的状态会影响输入输出之间的关系。也就是说，改变隐含层的权系数，可以改变整个多层神经网络的性能。

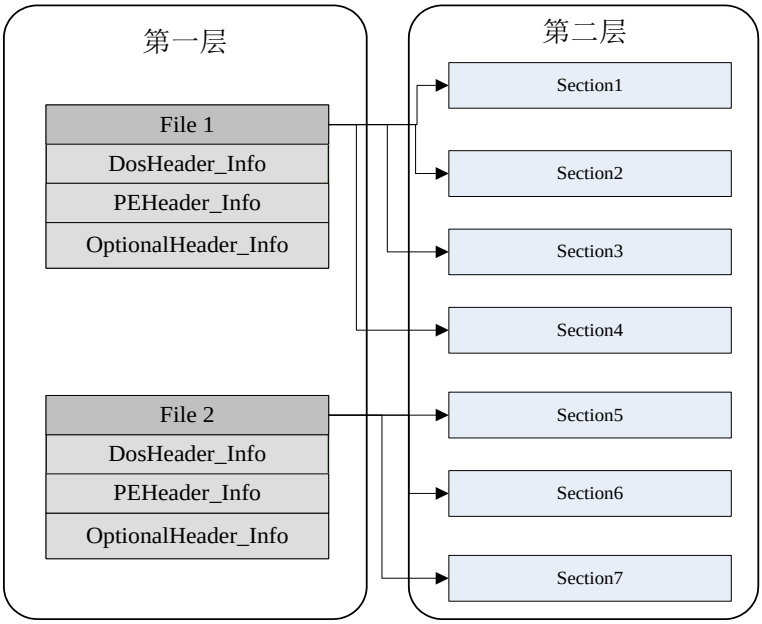
五、 基于 PE 静态信息的神经网络模型的建立

5.1 问题

由于 PE 文件的结构相对复杂，所以从 PE 文件中提取出的信息不能直接应用到神经网络进行训练，导致需要解决以下两个难点：

1) 神经网络的输入向量是一维的，并且每个 PE 文件的导入表，导出表可有可无，而且导入表和导出表均为非固定向量长度。这种结构实际上是多维的，

如下图所示：



所以首先必须对其进行降维处理。

2) BP 神经网络的输入向量必须取自（0，1）区间，而由于 PE 文件中的属性类型的多样性，如导入表的导入文件名、导入函数是字符串类型的，并且字符串是非定长的，而其他的一般属性可以直接通过读取出来的二进制数做直接转化，所以我们这里主要考虑如何将字符串归一化的问题。

5.2 解决方法

1) 多维结构的细粒化分析

对于如何解决 BP 神经网络的线性化要求和 PE 文件本身的非线性化数据之间的矛盾，肖新光先生最早提出来要应用基于变化粒度的分析方法，这是解决多维问题的一个非常好的思路，前面介绍了 PE 文件结构实际上是多维的，而神经网络训练的要求是一维输入、一维输出，那么我们如何利用变化粒度来解决多维的输入问题呢？

我们把基于变化粒度的方法描述如下:

①将属于同一层次（一对一结构）的属性单独训练，顺序为 小粒度 ->大粒度

②小粒度训练完毕后参加上一层的网络训练

③逐层训练，最终达到文件的顶层粒度，即与一个文件体的对应关系为 1-1。

2) 字符串数据的归一化

由于字符串是由一组 ASCII 码构成的序列，而每个单独的 ASCII 码是没有价值的，所以我们采用了概率统计的方法，通过对字符串进行分解统计和组合归一两个过程。巧妙地使每个字符串最终得到的概率值符合在区间（0，1）中，这样既做到了字符串的数字化，又满足了归一化要求。

3) 结果分析

首先，木马样本的质地本身对神经网络的训练结果构成了直接的影响，目前所采用存机取得样本的方法并不完全具有普遍性，如果换一种更为合理的采样方法，相信取得的效果会更为理想。

其次，对未知木马的报警率，结果是较为理想的。然而对于正常文件误报问题如果不能降低到 0.05%以下，是没有产品应用价值的。

最后，我们的挖掘模型尚有待改进和完善，同时我们对木马数据的特征潜力也充满信息。

六、 我们取得的进展

目前该技术已经被我们用于对未知文件的筛选（我们每月收到约 100000 个用户、网络嗅探和 Honey Pot 的文件上报），从实验结果看到，我们走的路是有价值的,并且在方向上基本是正确的。

我们对 PE 文件的信息的提取远没有开发完，目前我们只对 Dos Header, File Header, Optional Header, Section Header, Import Section 进行提取。而 Resource Section, Export Section 等重要的段还没有进行分析。

最后，目前该机制还没有与我们现有的自动特征挖掘体制结合。我们会在日后改善过程中，争取早日融入产品中去。

七、 一些思考

基于神经网络的静态评估的目标是对抗未知病毒的“整体”，并希望从中达到超过 50%实际报警率， 我们对此很有信心。

通过上面的方式可以看出，逃避这种体制并不困难。无论是基于特征码的检测还是 VM+启发式扫描，因为任何木马检测机制都不能对抗有目的的逃避。

最后，我们期冀 2005 年 5 月该技术会取得阶段性突破。应用于实际产品。

参考文献：

[1] Martin T.Hagan(美)等著. 戴葵等译. 神经网络设计 北京:机械工业出版社,2002

[2] Bjarne Stroustrup . C++ Programming Language(Special Edition) Pearson Education,2001

[3] 郑慧娆等著.数值计算方法 武汉：武汉大学生出版社，2002

