文章编号:1003-5850(2012)11-0020-04

基于分类的未知病毒检测方法研究

熊 俊
(湖南警察学院,长沙 410138)

摘 要:针对PE文件的静态信息,通过对未知病毒进行聚类分类分析,采用优化的初始聚类K-means算法,最终实现对病毒文件的相似度检测,无需运行PE文件即可判断其是否为病毒。该方法不仅克服了病毒特征码扫描无法识别未知病毒的缺点,而且相对于API序列检测方法免去了对文件进行脱壳等复杂操作,显著提高了检测速度。实验结果表明分类检测方法具有较好的准确性,有一定的应用价值。

关键词:信息安全,PE文件静态信息,未知病毒检测
中图分类号:TP309.5 **文献标识码**:A

Research of Unknown Virus Detection Based on Classification Method

XIONG Jun
(*Hunan Police Academy, Changsha*,410138,*China*)

Abstract:With PE file information as static characteristic, a classification method to detect unknown virus is proposed in this paper. In this paper, the K-means clustering algorithm based on the optimized initial cluster centers detects the similarity of the virus file . Without running the PE file, the classifier can determine whether it is virus or not. The method can overcome the shortage of virus feature scanning technology, which could not recognize unknown virus, and do not need for file shelling and other complex operations relative to the API sequence test methods, significantly improve the detection speed. Experiment results show that the detection method has better classification accuracy, so there is a certain practical value.

Key words:information security,PE file static information,unknown virus detection

随着网络技术的发展,当今病毒传播速度更快、破坏能力更强、使用更多的加壳及隐藏技术等特点,使防毒杀毒研究工作面临着巨大挑战[1]。目前主流杀毒软件使用的特征码扫描技术基本原理是:提取已知病毒样本中的一段二进制特征码,该特征码能唯一识别该类病毒,将此特征数据添加到病毒特征库中,在病毒检测时搜索病毒特征库查找是否存在匹配的病毒特征数据来检测是否病毒。该技术能快速和准确识别已知病毒,但对新型未知病毒无法识别或误报率高,同时大量已知病毒变种的传播无疑给检测过程带来很大阻碍[2]。如何对病毒进行高效且精确的分类,使分析过程自动识别过滤病毒变种并快速锁定新型病毒,对反病毒研究工作显得尤为重要。

目前较为流行的未知病毒检测方法是对文件行为进行监测和分析。基于多重朴素贝叶斯算法[3]和支持虚拟机[4]的未知病毒检测系统,均采用与数据挖掘技术相结合的方法,取得了很好的检测成功率。基于免疫的检测方法[5],采用了生物免疫系统的机制,具有较好的自学习机制。采用半增量贝叶斯算法构造分类器[6],在贝叶斯分类器基础上提高了学习知识的效率和分类效果。此外还有通过将未知病毒文件运行在虚拟机下检测病毒行为的方法,检测结果依赖于分析人员的专业水平,一般情况下能取得较高识别率,但检测的效率低下。基于API函数调用序列的数据挖掘技术无需运

* 收稿日期:2012-08-16,修回日期:2012-10-09
** 熊 俊,男,1974年生,讲师,硕士,研究方向:计算机信息安全。

行虚拟机，但病毒加壳、API 调用序列数量巨大等问题，给检测模型的实现带来很大难度。基于以上问题，本文提出了一种新的检测方法，将 PE 文件的重要静态信息作为特征码，通过进一步分析判断其是否为病毒文件。

本文首先分析了 PE 文件静态结构，选取文件重要信息作为特征向量，再通过数据挖掘分析对样本进行聚类，得到用于检测未知文件的分类器，最后通过实验对分类器的效果进行验证。

1　病毒检测模型及特征提取

1.1　本文使用的检测模型

系统检测模型结构如图 1 所示。系统由特征提取模块、构造分类器模块和病毒检测模块组成。

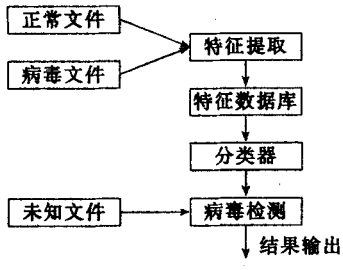

图 1　检测模型结构图

其工作原理如下：特征提取模块从训练集样本中抽取文件静态信息，经过一定的处理后得到统计数据并保存到数据库中；构造分类器模块使用初始聚类中心优化的 *K*-means 算法对测试集样本特征进行聚类分析，得到分类器和分类规则；病毒检测模块将测试集中的未知文件进行分类，根据所属类的性质输出是否为病毒，完成未知文件的检测。

1.2　病毒特征提取

PE(Portable Execute)文件是 Windows 操作系统中可移植的执行体，包括 exe、dll、ocx、sys、com 等，绝大多数的病毒文件都是 PE 格式的文件。MS-DOS 头部描述了文件的内容结构，紧接着的是模式残余程序是该应用程序加载后开始执行的位置。PE 文件标志和文件头包含了该文件的基本信息，如运行平台、区段个数等。PE 可选头部包含了可执行映像的重要信息，如程序入口点位置、初始堆栈大小，还有文件中一些表项的 RVA 与大小，如函数导入导出表等。紧接着的多个段分为段头和实体部分，每个段包含了各自的代码、数据及其他可执行信息，如导入函数表、模块导出函数表和重定位数据等。

常见的区段有 .text，.rdata，.data，.idata 等，本文重点关注的是函数导入区段 .idata，它记录了该 PE 文件所导入的相关系统 DLL 及其调用的函数。此外病毒还可能增加一个段来存放病毒代码，则会增加或修改段中某些关键域，如 Name、VirtualAddress、SizeOfRawData 等。由于 PE 文件的静态数据结构与装载到内存中的结构保持着很好的一致性，因此通过分析文件的静态信息来对 PE 文件进行分类判断具有可行性。

病毒文件在进行各种操作的时候与大部分 PE 文件相似，都是需要调用系统提供的接口和 API 函数，但它们常常通过调用一些较为特殊的系统 API 函数来达到其破坏目的，而这些函数在正常程序中很少使用。特殊的操作包括：对特定的注册表项进行读写操作(如添加开机启动项和加载项、注册为服务以达到自启动、修改隐藏文件夹选项来保护自己)；修改系统关键路径(如 system32)下的文件；对系统重要进程(如 explorer.exe、svchost.exe)注入钩子、创建远程线程、读写进程内存等。很多病毒文件还习惯于在达到目的后将自身删除或复制到较深的系统路径下，从而使人不易察觉和减少人的疑虑。因此检查文件中是否包含完成这些操作的 API 可以作为重要的判断特征。

对于使用显式链接来减少出现在导入函数表中 API，或对显式链接的 API 也进行加密的文件，仅仅使用 API 作为判断会导致准确率低下。相对数据来说，同种病毒的代码部分往往具有相似性。病毒代码中的字符串及一些数据可以随机产生，但由于同种病毒代码实现的是相同的功能，所以代码的内容本质上是一样的。在分析未知文件时，与特征库中已经存在的病毒信息相比较，其在文件结构、区段个数和大小等静态信息也较为相似，则很大可能是已知病毒的变种。因此把 PE 静态结构相似度作为判断依据。

由于很多病毒文件采用了加壳技术和加密技术，它们真正的代码往往在脱壳后才出现。对于无法脱壳的样本，仅读取代码入口附近的信息进行相似度比较来判断，就会使聚类变成壳的聚类，从而无法判断是否病毒文件。根据这一问题，通过求代码入口处上下一定地址范围内的数据区域大小、边界，及未知区域大小、边界等进行相似度比较。

综合以上，提取以下信息作为样本特征：文件导入表中调用的 DLL 文件个数和名称、每个 DLL 中调用的 API 函数个数及是否调用特殊 API 函数、代码中出现的特殊注册表路径和系统路径、代码相似度。表 1 为样本特征向量的一部分内容。

表 1 样本特征

类别	特 征
DLL 调用	Kernel32.dll，User32.dll，Shell32.dll，Advapi32.dll，GDI.dll，Imm32.dll，Ole32.dll，Msvcrt.dll，Comdlg32.dll，Oleaut32.dll，Setupapi.dll，Urlmoon.dll
API 调用	RegSetValue，CreateRemoteThread，WriteProcessMemory，CreateToolhelp32Snapshot，RtlAdjustPrivilege，CreateFile，WriteFile，CreateService，StartService，OpenProcess，SetWindowHook
关键注册表	HKLM\SOFTWARE\Microsoft\Windows\CurrentVersion\Run，HKCU\Software\Microsoft\Windows\CurrentVersion\Run ，HKLM\SOFTWARE\Microsoft\WindowsNT\CurrentVersion\Winlogon\[Shell]，HKLM\SOFTWARE\Microsoft\WindowsNT\CurrentVersion\Windows\[AppInit _ DLLs]
系统路径	C：\WINDOWS\system32，C：\WINDOWS\system32\drivers，C：\WINDOWS\system32\dllcache，C：\Documents and Settings\Administrator\Application Data
代码相似度	文件大小、是否签名、区段个数和名字、代码入口点所在节的位置、数据区域大小和边界

2 分类器的构造

2.1 基于 *K*-means 方法的分类

本文定义样本集 $U=\{U_1,U_2,U_3,\cdots,U_n\}$，每个样本用特征向量表示为 $U_i=\{X_{i1},,X_{i2},X_{i3},\cdots,X_{im}\}$，$X_{im}$ 表示样本 i 的第 m 个特征。

与传统 *K*-means 算法通过迭代过程把数据集划分为不同类别，使评价聚类性能的准则函数达到最优，从而使生成的每个聚类内部紧凑，类间相对独立，具有较好的聚类效果。

其算法步骤如下：

①对于样本集合 U，任意选取 k 个样本作为初始聚类中心；

②将样本集中的样本按照最小距离原则分配到最邻近聚类；

③使用每个聚类中的样本均值作为新的聚类中心，样本均值为 $\overline{X}_i=\frac{1}{|C_i|}\sum_{C_i}X$；

④重复②、③步骤直到不再发生变化；

⑤结束，得到 k 个聚类。

K-means 算法对大数据集能得到较好的聚类效果，类与类之间区别明显，但不足之处是结果依赖于初始聚类中心的选取，容易产生局部最优而影响结果，同时对聚类数 k 值的确定也没有一定计算方法，往往依靠主观判断。因此本文采用改进的 *K*-means 算法对训练样本集进行聚类，在初始化过程中动态地决定 k 的取值，同时获得较优的初始聚类中心集合。

2.2 对未知病毒的检测

通过对训练集进行聚类，得到 k 个聚类中心。定义样本 U_i,U_j 之间的相似度为：

$$S_{Ui,Uj}=1-\frac{\sum_{m=1}^{m}\frac{|X_{im}-X_{jm}|}{\mathrm{MAX}(X_{im},X_{jm})}}{n}$$

当两个样本相似度越高时，S 值越接近于 1，反之 S 越接近于 0。

对于待检测的未知样本 x，计算其与各个聚类的相似度，当相似度达到一定阈值则可以被判定属于该类，从而知道是否病毒文件。通过多次实验分析，当相似度达到 0.8 时就可判定所属的类，同时能取得较高的准确度和较低的误判率。当一个未知文件与所有的聚类比较均不匹配（即都没有达到相似阈值）时，将该样本作为一个新的聚类中心，并标记为特征库中未能识别的新类进入下一轮样本的检测。系统维护中对于所有未能识别的类，可结合其行为特征等方法判断是否为病毒，从而决定继续留在特征库中还是将其去除。为了维护特征库中信息的时效性，减少信息冗余导致匹配检测的时间消耗，还需要定期删除数据库中那些聚类个数很少、长时间没有新的文件特征与其匹配的条目。

3 实验结果分析

本文用于实验的样本空间共有样本总数 300 个，分为正常程序和病毒文件，分布如表 2 所示。其中正常程序是从新装的 Windows XP 系统中的 C 盘路径中选取，病毒文件从实验室研究的病毒库中获得，其中包括木马、后门程序、蠕虫。样本特征向量的获取借助反汇编逆向工具 IDA，该工具可以对文件的组成模块、系统调用等进行深入分析，如表 2 所示。

表 2 实验样本数据

	样本空间	训练集	测试集
正常文件	110	80	30
病毒文件	300	200	100
合计	450	320	130

实验的主要目的是验证上述模型对于未知病毒文件的分类和识别能力。对判定的结果是一个“是”和“否”的二分问题，该问题预测可产生 4 种结果，分别

为：①将正常程序判为正常，记为 TP；②将正常程序判为病毒，记为 TN；③将病毒文件判为正常，记为 FP；④将病毒文件判为病毒，记为 FN。准确率为测试样本被正确分类的概率；误报率为待分类的 175 正常文件被判为病毒的概率，如表 3 所示。

表 3 分类检测结果

类别名称	数量	改进 k_means 算法		k_means 算法		KNN 算法	
		准确率	误报率	准确率	误报率	准确率	误报率
木马	43	90.7%	2.9%	89.2%	3.4%	88.2%	3.8%
后门	31	93.2%	3.4%	78.7%	4.4%	88.7%	3.8%
蠕虫	23	91.6%	5.1%	86.3%	5.3%	89.3%	3.3%
正常文件	24	95.3%	3.1%	90.7%	6.1%	93.7%	4.1%

实验结论：

①使用初始聚类中心优化的 k_means 聚类算法相对于一般 k_means 聚类对未知文件分类和识别能力更好，准确率和误报率前者优于后者，效果令人满意；

②与基于其他算法的未知病毒检测方法比较，此模型判定一个未知文件所需的时间大约为 1 min 左右，而基于虚拟机的检测模型每个样本需要约 2 min ～3 min，可以看到检测效率大大提高；与其他需要脱壳操作的检测方法相比，其检测准确率和误报率相当，但实际上检测过程操作相对较少，且对于不断增加的新类具有可扩展性，因此在实际应用中更具有优势。

4 结 语

本文通过分析文件的静态信息，使用改进的 k-means 分类算法对计算机病毒进行检测和识别，具有较高的准确率和较低的误报率。由于采用的特征向量是文件调用 API 函数和文件本身的一些结构信息，它能很好地识别出已有病毒文件的变种。整个模型中提取文件特征向量、通过最大最小距离获取初始聚类中心、k-means 分类过程耗费较多开销。

该系统主要实现了病毒的静态检测和识别，对于已经运行的病毒无法起到查杀和恢复的作用，可以考虑与病毒查杀工具集成以及早发现病毒，降低其破坏性，减少不必要的损失。由于计算机中的文件只有很少一部分处在运行状态，如果将系统与文件下载实时监控技术相结合，在病毒文件还未运行的时候将其识别并清除，则对计算机的安全防护是一个很好的方案。

参考文献：

[1] Peter S. The Art of Computer Virus Research and Defense [M]. Massachusetts: Addison Wesley Professional, 2005.

[2] Robert M, Ido G. Detection of Unknown Computer Worms Activity Based on Computer Behavior using Data Mining [C] // Proceedings of the 2007 IEEE Symposium on CISDA, 2007: 169-177.

[3] 张波云，殷建平，蒿敬波，等. 基于多重朴素贝叶斯算法的未知病毒检测[J]. 计算机工程，2006，32(10)：18-21.

[4] 张波云，殷建平，蒿敬波. 基于 SVM 的计算机病毒检测系统[J]. 计算机工程与科学，2007，29(9)：19-22.

[5] 陈 恒，刘晓洁，宋 程，等. 一种基于免疫的计算机病毒检测方法[J]. 计算机应用研究，2005，22(9)：111-114.

[6] 赖英旭，李 征. 未知病毒检测技术的研究[J]. 计算机科学，2006，33(8)：300-301.

(上接第 19 页)

```
while ((res = pcap_next_ex( adhandle, &header, &pkt_data)) >= 0){
    if(res == 0) //超时时间到
      continue;
    printf("len:%d\n", header->len);
}
```

4 总 结

MAC 地址泛洪攻击是一种常见的攻击技术，主要通过发送伪造源 MAC 地址数据帧达到攻击目的。攻击会使交换机中流量过大，造成网速变慢、数据帧丢失甚至完全瘫痪的严重后果。针对这种攻击手段，一般可以采用以下几种方法进行预防：

①划分 VLAN，分割广播域，尽量使攻击的影响范围缩小。

②限定 MAC 地址表中一个端口映射的 MAC 地址数量。即在交换机上配置端口安全，让端口和 MAC 地址绑定或限制一个端口连接的计算机数目。

③对于交换式以太网络，采用二层访问控制和认证协议 IEEE802.1x 进行用户接入控制[2]。

参考文献：

[1] The WinPcap Team, CoffeeCat 译. WinPcap 中文技术文档[J/OL]. http://www.ferrisxu.com/WinPcap/html/index.html, 2007-10-08.

[2] leadlxx 博客. MAC 地址泛洪攻击[J/OL]. http://leadlxx.blog.51cto.com/153568/317535, 2010-06-07.

[3] 赵建勋. 基于 WinPcap 网络数据包捕获实现[J]. 西安文理学院学报(自然科学版)，2010，14(4)：55-58.

基于分类的未知病毒检测方法研究

作者： 熊俊， XIONG Jun
作者单位： 湖南警察学院,长沙,410138
刊名： 电脑开发与应用
英文刊名： Computer Development & Applications
年，卷(期)： 2012,25(11)

参考文献(6条)

1. Peter S The Art of Computer Virus Research and Defense 2005

2. Robert M;Ido G Detection of Unknown Computer Worms Activity Based on Computer Behavior using Data Mining 2007

3. 张波云;殷建平;蒿敬波 基于多重朴素贝叶斯算法的未知病毒检测[期刊论文]-计算机工程 2006(10)

4. 张波云;殷建平;蒿敬波 基于SVM的计算机病毒检测系统[期刊论文]-计算机工程与科学 2007(09)

5. 陈恒;刘晓洁;宋程 一种基于免疫的计算机病毒检测方法[期刊论文]-计算机应用研究 2005(09)

6. 赖英旭;李征 未知病毒检测技术的研究 2006(08)