基于改进的 K⁻最近邻算法的病毒检测方法

谢金晶,张艺濒
(武汉大学 计算机学院 湖北 武汉 430072)

摘 要:由于计算机病毒检测的不可判定性,提出了一种基于改进的 K^-最近邻检测方法来实现对计算机病毒的近似判别。此方法成功地克服了现有的特征码扫描技术只能检测已知病毒的缺点。首先改进了原始的 K^-最近邻检测方法,使其更适合于对计算机病毒进行预测。并在此检测方法上,设计了一个病毒检测系统。此系统既可查杀已知病毒,也可分析评判可疑程序,诊断出被感染病毒以及病毒类型。

关键词: K^-最近邻算法;计算机病毒;病毒检测;Internet
中图分类号:TP302.1 **文献标识码**:B **文章编号**:1004-373X(2007)03-051-03

Computer Virus Detection Based on Improved K⁻ Nearest Neighbor Algorithm

XIE Jinjing,ZHANG Yibin
(Computer College,Wuhan University,Wuhan,430072,China)

Abstract:Because precise determination of a computer virus is undecidable,a method based on improved K^- nearest neighbor to detect computer virus approximately is presented in this paper. It can successfully overcome the disadvantage of normal virus scanner,which can only detect known virus. First improved the primal K^- nearest neighbor algorithm, which can make it more suitable for computer virus detection. And a virus detect system based on this detect method is also designed. This system can detect known computer virus and can analysis and judge shadiness program,diagnose which kind of computer virus is infected.

Keywords: K^- nearest neighbor algorithm;computer virus;virus detection;Internel

1 引 言

随着 Internet 的迅速发展,计算机病毒问题也日益严重。新生病毒数量越来越多,破坏性、隐藏性、智能性也越来越好,给人们造成的损失也越来越大。检测防范病毒刻不容缓。当前的计算机病毒检测技术主要是以特征码扫描技术为主。但此检测方法只能检测出已知病毒,而对未知病毒、多态性病毒及隐蔽性病毒则无能为力。针对此问题,本文提出一种基于改进的 K^-最近邻算法的病毒检测技术:用改进的 K^-最近邻算法对可疑程序代码特征进行分析,并利用病毒和正常程序的代码特征的差异性进行分类,从而检测出未知病毒。基于此方法,本文设计了一个计算机病毒检测系统并介绍了他的工作原理。此系统既可查杀已知病毒,也可分析评判可疑程序,诊断出被感染病毒以及病毒类型。

2 基于改进的 K⁻最近邻算法的病毒检测系统

2.1 系统结构分析

如图 1 所示,当文件由 Internet 进入检测系统后,首先由病毒防火墙进行拦截。病毒防火墙可检测并查杀出当前已知病毒。若未检测出病毒,则将其分别送入服务器和特征提取器中。在特征提取器中用相应的特征提取程序对文件的代码进行特征提取。并在病毒检测服务器中用改进的 K^-最近邻算法对代码特征进行分析,判别是否为病毒。若是病毒的话,则在病毒分类器中利用改进的 K^-最近邻算法的运算结果诊断出病毒所属种类,并将处理策略反馈给服务器及病毒防火墙。

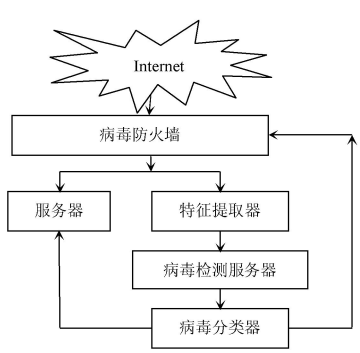


图 1 系统结构框图

2.2 病毒特征提取

病毒和正常程序的区别体现在很多方面,如一般的应用程序在最初的指令是检查命令行输入有无参数项,清屏和保存原来屏幕显示等,而病毒程序则从来不会这样做,他通常最初的指令是远距离跳转、搬移代码、直接写盘操作、解码指令或搜索某路径下的可执行程序等相关操作指令序列。故可通过扫描程序代码中可疑功能操作对应的

收稿日期:2006-08-16
基金项目:湖北省自然科学基金(2005ABA238)资助

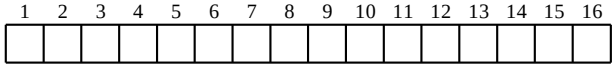
指令序列,来获取程序相应的代码特征。

可选取的病毒属性如表 1 所示。

表 1　可选取的病毒属性

序号	可疑功能操作指令序列
1	可疑的内存分配操作,程序使用可疑方式进行内存申请和分配操作
2	返回程序入口,在完成对原程序入口处开始的代码修改之后重新指向修改之前的程序入口
3	直接写盘操作,程序不通过常规的 DOS 功能调用而进行直接写盘操作
4	变化的程序入口程序被蓄意设计成可编入宿主程序的任何部分
5	具有可疑的文件操作功能,有进行感染的可疑操作
6	无效操作指令,仅仅用来实现加密变换或逃避扫描检查的代码序列
7	可疑的跳转结构,使用了连续或间接的跳转指令
8	非正常堆栈
9	程序截获其他软件的加载和装入
10	内存驻留程序
11	在内存中搬移或改写程序的代码序列
12	程序以可疑的方式进行重定向操作
13	不合逻辑的错误的时间标记
14	未公开的中断 DOS 功能调用
15	EXE/ COM 辨认程序
16	解码指令序列

将此 16 个属性按其序列号对应于两字节中,如下:



则当程序代码具有相应位所对应的特征时,此位就被填上 1,否则就添 0。故可得到文件特征向量为:

$$x = < t_1(x), t_2(x), \cdots, t_{16}(x) >,$$

$$t_i(x) = \{0,1\}, 1 \leq i \leq 16$$

2.3　病毒检测

病毒检测服务器的功能为利用改进的 K-最近邻算法检测可疑程序是否为病毒。此处针对病毒检测的特征,对 K-最近邻算法做了相应改进,使其能更准确快速地对病毒进行检测。

2.3.1　改进的 K-最近邻算法

K-最近邻算法(KNN)是一种基于类比的学习方法。KNN 算法的关键技术是搜索 N 维模式空间找出最接近未知样本的 K 个训练样本,未知样本被分配到 K 个最近邻者中最公共的类,其近邻性用欧几里德距离定义。用最近邻方法进行预测的理由是基于假定:一个实例的分类与在欧氏空间中他附近的实例的分类相似。

但原始的 K-最近邻算法具有分类不够准确、计算量大等缺点,故从以下几点进行改进,使其更适合于对计算机病毒进行预测。

(1) 对特征向量中各维属性加权

假设所有实例都惟一对应于 N 维空间中一个点 x,则由上文可知其特征向量为 $x = < t_1(x), t_2(x), \cdots, t_{16}(x) >$,故任意两个实例 x_i, x_j 之间的距离为:

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{16}(t_k(x_i) - t_k(x_j))^2} \quad (1)$$

由式(1)可看出实例的各维属性对实例之间的距离的影响是均等的,但是这对病毒的检测并无好处。因为显然各属性对应的安全和可疑级别并不相同。则为了使 $\mathrm{dist}(x_i, x_j)$ 能更逼真地反映出两实例之间的距离,根据病毒可能使用和具备的特点而对各维属性 $t_i(x)$ 授予不同的加权值 w_i。则将式(1)改为:

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{k=1}^{16} w_k(t_k(x_i) - t_k(x_j))^2} \quad (2)$$

(2) 对 K 个近邻按距离加权

在本文的最近邻学习中,离散目标函数为: $f: R^{16} \rightarrow V$。其中 $V = \{v_1, v_2\}$, v_1 表示为病毒, v_2 表示为正常程序。由于 K 最近邻算法一定会寻找到 K 个与其距离最近的点,则若点分布稀疏时,会覆盖较大的范围,则有可能受到样本空间中孤立点的影响。考虑到实际中实例受其近邻的影响是和距离成反比的,则对 K-最近邻算法按距离加权,则有:

$$f(x_q) \leftarrow \arg\max \sum_{i=1}^{k} w_i \delta(v, f(x_i))$$

其中 w_i 为权重,一般取每个近邻与查询点 x_p 的距离的平方的倒数,即:

$$w_i = \frac{1}{\mathrm{dist}(x_p, x_i)^2}$$

(3) 采用剪辑近邻法对样本集进行剪辑

K-最近邻算法的一个严重不足是会导致大量的计算量,因为他推迟所有的处理直到接收到一个新的查询。则采用剪辑近邻法,利用现有样本集对其自身进行剪辑,将不同类别交界处的样本以适当方式筛选,可以实现既减少样本数又提高正确识别率的双重目的。

剪辑的过程是:将样本集 K_N 分成两个互相独立的子集:test 集 K_T 和 reference 集 K_R。首先对 K_T 中每一个 X_i 在 K_R 中找到其最近邻的样本 $Y_i(X_i)$。如果 Y_i 与 X_i 不属于同一类别,则将 X_i 从 K_T 中删除,最后得到一个剪辑的样本集 K_{TE}(剪辑样本集),以取代原样本集,对待识别样本进行分类。

则由此可得到改进的 K-最近邻算法,可用此方法实现对未知病毒的有效检测。

2.3.2　病毒检测服务器

当采用此病毒检测系统对病毒进行检测前,需要先对其进行训练,得到一定数量训练集。则首先需要搜集大量的病毒程序及正常程序。将其按照上文中的方法提取得到特征向量,并通过改进的 K-最近邻算法对其进行分类,最后将分类结果加入训练集中。

对可疑程序的检测,即利用改进的 K-最近邻算法判别其是否为病毒。则若已知一训练集,有 N 个训练样本,对一个可疑程序 s_p,他的 K 个最近邻中有 n_1 个为病毒,

n_2 个为正常程序。则若 $n_1 > n_2$,可疑程序 s_p 为病毒,若 $n_1 < n_2$,可疑程序 s_p 为正常程序。由于 K 一般取奇数,则不可能出现 $n_1 = n_2$ 的情况。

2.4 病毒分类

由改进的 K-最近邻算法,可得到可疑程序 s_p 的 K 个最近邻,并通过比较 n_1,n_2 的大小可得知其是否为病毒。假设 $n_i(i = 0,1)$ 较大,则 s_p 与 n_i 个近邻同类。若其均为病毒,则可通过 n_i 个近邻的特征向量的计算得到可疑程序 s_p 的特征向量,从而判断其所属病毒类别。考虑到近邻与 s_p 的距离对 s_p 仍有影响,则仍采用上文中的权重 $w_i = \frac{1}{\mathrm{dist}(x_p, x_i)^2}$ 对 n_i 个近邻的特征向量进行加权,则得到可疑程序 s_p 的特征向量计算式为:

$$x_p = \sum_{j}^{n_i} x_j$$

即:

$$\begin{bmatrix} t_{p,1}(x) \\ t_{p,2}(x) \\ \cdots \\ t_{p,16}(x) \end{bmatrix} = w_1 \begin{bmatrix} t_{1,1}(x) \\ t_{1,2}(x) \\ \cdots \\ t_{1,16}(x) \end{bmatrix} + w_2 \begin{bmatrix} t_{2,1}(x) \\ t_{2,2}(x) \\ \cdots \\ t_{2,16}(x) \end{bmatrix} + \cdots + w_n \begin{bmatrix} t_{ni,1}(x) \\ t_{ni,2}(x) \\ \cdots \\ t_{ni,16}(x) \end{bmatrix}$$

由于病毒的特征向量中 $t_{i,j}(x) = \{0,1\}$,而由上式加权计算后得到的 $t_{i,j}(x)$ 很有可能为一小数,则设定一个阈值 $T = ni/2$,若 $t_{p,i}(x) > T$,则令 $t_{p,i}(x) = 1$,若 $t_{p,i}(x) \leq T$,令 $t_{p,i}(x) = 0$。当得到由 0,1 表征的特征向量后,可根据特征向量及特征向量每项对应病毒特征判断病毒类型进而提出相应的处理方法。

3 结 语

由于现有的计算机病毒检测方法:特征代码法仅能检测出已知病毒,而对未知病毒则无能为力,故本文针对 K-最近邻算法易受孤立噪声影响及计算量大等缺点加以改进,并基于此改进算法提出了一种病毒检测系统。此系统既可查杀已知病毒,又可准确高效地检测出未知病毒,并可由未知病毒的特征向量诊断出病毒所属类型,为处理未知病毒提供依据。

参 考 文 献

[1] Prabhat K Singh, Arun Lakhotia. Analysis and Detection of Computer Viruses and Worms: An Annotated Bibliography. ACM SIGPLAN Notices, Column: Technical Correspondence Table of Contents, 2002, 37(2): 29-35.

[2] Chenyi Xia, Wynne Hsu, Mong Li Lee. ERkNN: Efficient Reverse K-nearest Neighbors Retrieval with Local kNN-distance Estimation. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005: 533-540.

[3] Diomidis Spinellis. Reliable Identification of Bounded-Length Vinuses Is NP-Complete [J]. IEEE Transaction on Information Theory, 2003, 49(1): 280-284.

[4] Gerald J Tesauro, Jeffrey O Kephart, Gregory B Sorkin. Neural Networks for Computer Virus Recognition[J]. IEEE Expert, 1996; (8): 5-6.

[5] 张波云,殷建平,张鼎兴,等. 基于 K-最近邻算法的未知病毒检测[J]. 计算机工程与应用, 2005, 41(6): 7-10.

[6] 郭晨,梁家荣,梁美莲. 基于 BP 神经网络的病毒检测方法[J]. 计算机工程, 2005, 31(2): 152-153, 156.

[7] 傅建明,彭国军,张焕国. 计算机病毒分析与对抗[M]. 武汉:武汉大学出版社, 2004.

(上接第 50 页)

软件解调判定流程如图 5 所示。

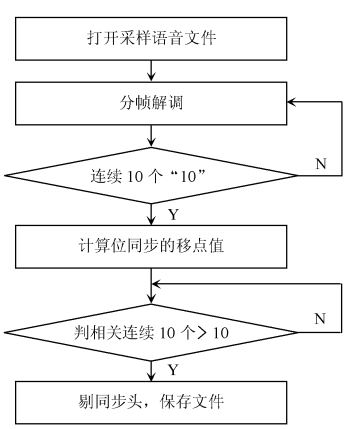


图 5 软件判定流程图

采用 Matlab 编程实现,利用 MathTools MIDEVA 4.5 把其转换成 *.DLL 文件,嵌入到 VB 语言编写的主程序中,运行良好。

5 结 语

以上只考虑在信干比较大的情况下的语音同步判定,已经在系统中应用,效果较好。在有较大干扰的情况下,语音如何同步,需要做进一步研究。

参 考 文 献

[1] 李素芝,万建伟. 时域离散信号处理[M]. 长沙:国防科技大学出版社, 1994.

[2] Shinsuke Hara, Attapol Wannasarnmaytha, Yuuji Tsuchida, *et al*. A Novel FSK Demodulation Method Using Short-time DFT Analysis for LEO Satellite Communication Systems. TEEE Trans. on Vehicular Technology, 1997, 46: 625-632.

[3] 樊昌信,詹道庸,徐炳祥,等. 通信原理[M]. 4 版. 北京:国防工业出版社, 1995.