基于N-Gram的计算机病毒特征码自动提取的改进方法

杨 燕[1] 蒋国平[2]

(南京邮电大学计算机学院 南京 210003)[1] (南京邮电大学自动化学院 南京 210003)[2]

摘 要 随着计算机技术的发展和普及,计算机病毒带来的危害日趋严重。传统N-Gram算法难以提取不同长度的特征,导致有效特征缺失,并产生庞大的特征集合,造成空间的浪费。针对这些问题,提出一种改进的基于N-Gram的特征码自动提取方法。该方法在原有N-Gram特征提取算法的基础上引入变长N-Gram特征,提取不同长度的有效特征,生成不定长病毒特征码。综合考虑特征频率的相关性,利用特征浓度对N-Gram特征进行有向筛选,生成数据字典,节省存储空间。实验结果表明,与单纯使用定长N-Gram的算法相比,该方法能有效降低特征码自动提取的误报率。

关键词 N-Gram,病毒特征码,特征浓度,数据字典

中图法分类号 TP309.5 **文献标识码** A

Improved Method of Computer Virus Signature Automatic Extraction Based on N-Gram

YANG Yan[1] JIANG Guo-ping[2]

(School of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[1]

(School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[2]

Abstract With the rapid development of computer technology, security threats brought by computer virus have become more and more serious. The traditional N-Gram algorithm is difficult to capture bytes of different length, leading to the lack of effective signature and the geheration of huge signature sets, and creating a waste of storage space. Instead of using fixed-length N-Gram feature that the traditional way dose, an improved computer virus signature automatic extraction algorithm based on variable-length N-Gram was proposed to solve these problems. It extracts the effective signature to generate variable-length virus signature. Taking the correlation of signature frequency into account, the algorithm uses signature concentration to extract the N-Gram feature of malware samples and generates a data dictionary to save the storage space. In the experiment results, compared with the traditional algorithm which uses fixed-length N-Gram feature, the proposed method can effectively decrease the false rate of signature extraction.

Keywords N-Gram, Virus signature, Signature concentration, Data dictionary

1 引言

随着计算机技术在全球的蓬勃发展,计算机病毒带来的危害日趋严重。研究计算机反病毒技术并有效查杀计算机病毒,是当前信息安全技术研究领域的重中之重。因此,反病毒技术应运而生,诸如特征码扫描技术、行为检测技术、启发式扫描技术及虚拟机技术。行为检测技术通过分析病毒与正常程序的行为之间的差异性来检测病毒,该方法可以检测未知病毒,但误报率较高。启发式扫描技术通过对比指令代码,实现病毒检测工作;该方法能够提高扫描效率,检测未知病毒,但是对经过加壳保护处理后的计算机病毒却无能为力。虚拟机技术是将病毒程序运行在虚拟机上并进行监控的技术;该方法能有效查杀加密和多态病毒,但存在模拟范围有限、速度缓慢、资源消耗量大等不足。当前特征码扫描法是检测已知病毒最简单且最有效的方法,也是目前计算机反病毒软件中应用最广、最基本的病毒检测技术,其基本原理是在内存中划分一部分空间,将电脑中流过内存的数据与反病毒软件自身所带的病毒库的特征码相比较,以判断其是否为病毒。如今,业界主流的杀毒软件(如卡巴斯基、McAfee、360安全卫士)都主要采用特征码扫描法对病毒进行检测。

一个新病毒在互联网上出现之后,尽快提取其特征码并将其更新至用户所安装的杀毒软件特征库中,对于遏制该病毒的传播至关重要。然而,现今的特征码提取方法仍停留在人工手动提取阶段。为此,有学者对计算机病毒特征码自动提取技术进行了研究,并相应地提出了很多有效算法,包括基于概率有限状态自动机的特征码提取算法、基于生物遗传的特征码提取算法、基于Rabin Fingerprints的特征码提取算法、基于最长公共子串的特征码提取算法、基于N-Gram的特征码提取等。例如,Yegneswaran等人[1]对聚类形成的类簇进行归纳处理,将其转化为可被有限状态自动机所识别的正则语言,并自动提取特征码,但难以提取大量的变形特征码。Lee等人[2]利用遗传算法提取特征,采用统计方法提取出特征码。Kijewski等人[3]根据递推来比较模式串与正文子串的哈希函数值以提取特征,降低了时间复杂度和空间复杂度。

杨 燕(1990-),女,硕士生,主要研究方向为网络安全技术与应用,E-mail:1094164156@qq.com;蒋国平(1966-),男,博士,教授,博士生导师,主要研究方向为混沌系统、复杂动态网络、信息安全。

Kreibich 等人[4]的算法简单,但时间复杂度大,占用了大量空间。尽管上述算法相对于人工手动技术可以更快地查杀病毒,但是在提取特征的过程中,它们面对潜在的有效特征时很难进行准确提取。采用 N-Gram 算法可以有效地解决难以提取潜在特征的问题,从而降低误报率[5]。1994 年,Kephart 首次将 N-Gram 模型应用到计算机病毒检测领域[6]。张福勇等人[7]使用 N-Gram 算法提取特征,但存在特征冗余、特征集合庞大的问题,增加了寻找区分恶意软件和正常软件的特征的难度。如何选择有效特征,对于提高病毒检测的性能至关重要[8-9]。很多人提出了有效特征的选择方法,并在实际中取得较好结果,但其中仍有不完善的方面。例如:在筛选有效特征时采用随机刷选,无向导性,导致产生庞大的特征集合,生成效率低下;没有充分利用多个特征之间的相关性;特征码的提取长度受限,不能推广到更广泛的环境中。

针对上述不足,本文提出一种改进的 N-Gram 特征的提取算法,其利用特征浓度选取变长的 N-Gram 特征,通过对比匹配样本程序和特征库生成特征码,并借助 N-Gram 统计语言模型刷选特征码。

本文将计算机病毒特征码自动提取技术作为研究主题,第 2 节简要描述所提出的特征码提取模型的结构;3.1 节概述 N-Gram 特征提取算法,3.2 节提出变长 N-Gram 特征提取算法,3.3 节详细阐述病毒特征码生成;第 4 节通过分析实验结果,得出主要结论。

2　特征码提取模型的结构

本文实现了一种改进的 N-Gram 特征提取方法,其主要分为两个部分。1)训练部分:首先选取若干个病毒程序和合法程序,提取样本程序的变长 N-Gram 作为特征,根据特征的有向选择计算特征浓度值 S_i,通过 S_i 降序生成数据字典,选取数据字典中序号 SN 大于 S 的有效特征作为最终生成的病毒 N-Gram 特征库,以供待测程序匹配。2)生成部分:首先将病毒程序转化为十六进制格式,将待测病毒样本与病毒 N-Gram 特征库库进行匹配,当待测病毒样本字节流包含的病毒特征库中的特征的数目 L 大于等于某一阈值 C 时,提取该字节流作为病毒样本的候选特征码。借助 N-Gram 统计语言模型,计算候选特征码出现的概率,选择概率最小的候选特征码作为最能代表该病毒程序的特征码,从而实现计算机病毒特征码的自动提取工作。该模型的训练流程和生成流程分别如图 1 和图 2 所示。

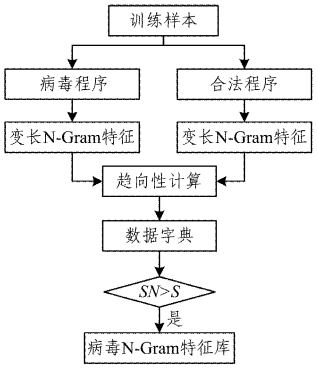

图 1　模型的训练流程

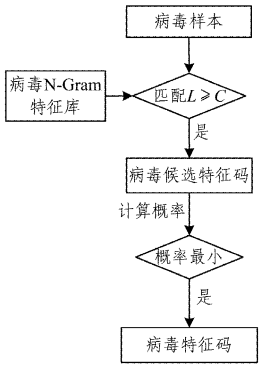

图 2　模型的生成流程

针对本文提取计算机病毒特征码的算法,有如下两个评价指标[10]。

成功率(SR):成功提取出的特征码的病毒数目在待测病毒样本中所占的比例。

误报率(FPR):根据提取出的病毒特征码检测出的合法程序数目在待测合法程序样本中所占的比例。

3　特征提取方法

3.1　N-Gram 算法

N-Gram 模型是一种用于预测一个序列中下一项的概率的语言模型,它类似于 $n-1$ 阶的马尔可夫链,可用于概率论、计算机语言、通信理论等很多领域[11-12]。在恶意代码的检测中,N-Gram 的思想是将程序文件视为一个字节流,首先提取程序文件的字节序列。使用大小为 N 的滑动窗口进行操作,得到长度为 N 的字节序列,即 N-Gram 特征。当提取的 N-Gram 特征为 3 字节时,滑动窗口方法的示意图如图 3 所示。

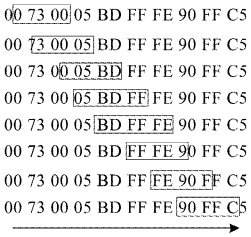

图 3　滑动窗口方法

按照上述思路描述,训练部分的特征选取算法总结如下:

假设 T_v 是训练样本中所有病毒程序的 N-Gram 特征总数;T_p 是训练样本中所有合法程序的 N-Gram 特征总数;C_v 是训练样本中所有病毒程序的个数;C_p 训练样本中所有合法程序的个数;F_v^i 是特征 i 在病毒程序中出现的次数;F_p^i 是特征 i 在合法程序中出现的次数;C_v^i 是包含特征 i 的病毒程序个数;C_p^i 是包含特征 i 的合法程序个数;G_j 表示第 j 个类别,此处一共有 2 类,即病毒程序和合法程序,其中 $j\in\{0,1\}$。

(1)将所有样本文件转成十六进制格式;

(2)初始化 $T_v,T_p,C_v,C_p,F_v^i,F_p^i,C_v^i,C_p^i$ 和 G_j 为 0;

(3)选择一个病毒程序,并初始化一个标志数组 Flag[i]=0,C_v++,G_j=1;

(4)使用大小为 3 的滑动窗口,收集窗口内的 N-Gram 特征,记作索引 i,F_v^i++,Flag[i]=1,标志 N-Gram 特征已在该程序中出现过一次;

(5)窗口向前滑动 1 个字节,并返回步骤(3),直到该病毒程序结束;

(6)若 $G_j=1$ 且 Flag[i]=1,则 C_v^i++,表明包含该 N-Gram 特征的病毒程序数目;

(7)返回步骤(3),直到统计完所有病毒程序并选择一个合法程序,初始化一个标志数组 Flag[i]=0,C_p++;

(8)使用大小为 3 的滑动窗口,收集窗口内的 N-Gram 特征,记作索引 i,F_p^i++,Flag[i]=1,标志 N-Gram 特征已在该程序中出现过一次;

(9)窗口向前滑动 1 个字节,并返回步骤(8),直到该合法程序结束;

(10)若 $G_j=0$ 且 Flag[i]=1,则 C_p^i++,表明包含该 N-Gram 特征的合法程序的数目;

(11)返回步骤(8),直到统计完所有合法程序。

但是,N-Gram 也有其自身的缺陷:N-Gram 很难提取到不同长度的字节序列,如果一个程序样本中含有多个有效、不等长的特征,或者长度不是 N 的整数倍时,则会产生边缘无匹配,从而无法同时提取这些特征,造成有效特征片段丢失;同时,N-Gram 会产生庞大的特征集合,需要相当大的空间来存储,造成空间浪费[13-14]。本文在提取特征时对特征频率进行趋向性计算,生成数据字典,减少存储空间。

对于一个特征,如果该特征出现的次数在所有病毒 N-Gram 特征中所占比例越大,在所有合法 N-Gram 特征中所占比例越小,则该特征被选为有效性特征的概率就越大。同理,若在训练样本中包含该特征的病毒程序数目在所有病毒程序数目中所占的比例越大,在训练样本中包含该特征的合法程序数目在所有合法程序数目中所占的比例越小,则该特征被选为有效特征的概率越大。

因此,基于上述条件,本文提出一个有效的 N-Gram 特征有向选择公式:

$$S_i=\log(1+\frac{D_p^i}{D_v^i}) \tag{1}$$

其中,$D_v^i=\frac{C_v^i\times F_v^i}{T_v\times C_v}$,表示 N-Gram 特征出现在病毒程序中的概率;$D_p^i=\frac{C_p^i\times F_p^i}{T_p\times C_p}$,表示 N-Gram 特征出现在合法程序中的概率。

3.2 变长 N-Gram 算法

变长 N-Gram 是一串有意义的连续字节序列,与 N-Gram 的区别在于变长 N-Gram 的长度是可变的,避免了一个有意义的字节序列被拆开。提取有效的变长 N-Gram 特征时必须选择合适的特征长度。使用 3-Gram 特征和 4-Gram 特征对区分病毒程序和合法程序的效果更好,本文提取的 N-Gram 字节为 3 字节[13],通过统计与合并 3-Gram 特征,生成变长 N-Gram 特征。

改进后的特征提取算法如图 4 所示。具体步骤如下:

(1)将所有程序样本转化为十六进制格式。

(2)选择一个程序样本,提取该样本中所有的 3-Gram 特征,将特征记为 G_i,同时统计其出现的频率并记为 F_i。

(3)如果在当前程序样本中 G_i 之前存在一个 3-Gram(G_{i-1}),则记入 gram 关联矩阵 $A(G_{i-1},G_i,M)$,其中 M 表示 G_{i-1} 和 G_i 同时出现的频率。

(4)如果 M 大于阈值 T,则合并两个特征成为新特征并将它加入特征集合中,从特征集合中删除原特征,其中 T 代表合并阈值比例。如果两个 N-Gram 特征连续出现的频度达到这两个 N-Gram 特征各自出现频率的规定比例 T,则认为这两个 N-Gram 特征应该合并为一个 N-Gram 特征。经实验多次验证,T 取值为 75%时效果比较理想。

(5)计算特征浓度 S_i,对变长 N-Gram 特征进行有向选择,通过 S_i 降序生成数据字典。

(6)选取数据字典中序号 SN 大于 S 的有效特征作为最终生成的病毒 N-Gram 特征库。

(7)返回步骤(2),直到统计完所有程序样本。

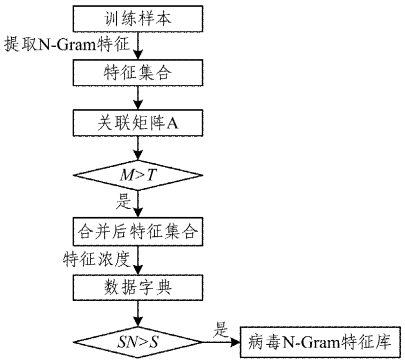

图 4 改进的 N-Gram 特征提取算法的流程

训练样本中部分变长 N-Gram 特征信息如表 1 所列。

表 1 部分变长 N-Gram 特征信息

N-Gram	F_v^i	F_p^i	C_v^i	C_p^i	S_i
726FC5	49956	1136	146	85	1.379529
534F92	4563	501	198	103	0.800887
085A37FF	12947	946	133	116	0.761198
50ED61	7370	549	104	94	0.741559
92F75D	6631	1521	152	121	0.425676
5D84455C	2407	918	118	153	0.254571

表 1 中特征浓度 S_i 值的大小与特征的有效性成正比,因此本文把特征浓度 S_i 值较大的特征组合成病毒特征库。按特征浓度 S_i 值降序排列,生成数据字典,从而节省存储空间,降低特征冗余。数据字典是一张参照表,根据数据字典可以判断哪些特征作为有效特征比较合适。从数据字典中可以选择出前 S 个有效 N-Gram 特征作为最终生成的病毒特征码的 N-Gram 特征库。其中,S 是调节参数,通过不断实验可以得出一个合适的阈值。一个合适的阈值不仅能够保持病毒特征库的数目最小,而且可以保障病毒特征库的特征最能够代表病毒程序。

3.3 病毒特征码的生成

如果计算机病毒特征码长度过短,那么生成的特征码不能包含足够多的信息,就会造成候选特征码的数量庞大,包含过多的无效特征码,从而降低算法性能并减弱效果。反之,如果特征码长度过长,那么对匹配长度则具有更高的要求,时间和空间复杂度随之增加,从而使算法可执行性降低。综合考虑,在保持唯一性的前提下,应尽量使特征码长度更短。

计算机病毒特征码的具体生成步骤如下:

(1)将待测程序转化为十六进制格式;

(2)待测程序与计算机病毒特征库进行匹配;

(3)从第一个发生匹配的位置开始,采用滑动窗口的方式向后进行匹配比较,一直匹配前进,直到发生间断为止;

(4)统计该段字节流中包含病毒特征库中的特征的数目;

(5)如果数目 L 超过某个阈值 C,则将该段字节流加入计算机病毒特征码候选库,将其作为一个候选特征码;

(6)继续与计算机病毒特征库进行匹配,返回步骤(3),直到提取出该病毒样本的所有候选特征码。

综合考虑,在保持唯一性的前提下,应当尽量使得计算机病毒特征码短小精悍,因此本文规定 C 取 3。不定长计算机病毒候选特征码的生成过程如图 5 所示。

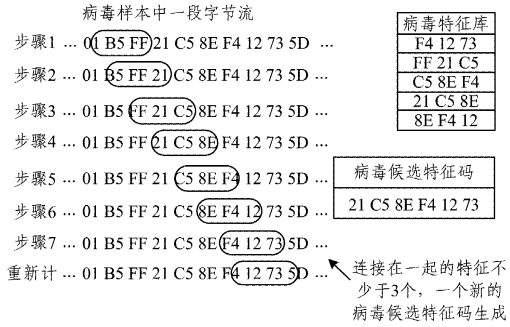

图 5　计算机病毒候选特征码的生成

如果计算机病毒特征码候选库中包含的特征码数目较多,则需要再次选择特征码,使被选择到的特征码最能够表示该病毒,将该病毒和其他病毒或者合法程序区别开。本文借助 N-Gram 统计语言模型对病毒特征码进行选择,提取最合适的特征码。

N-Gram 是一种统计语言模型,也是常用的语言识别模型,用于统计句子出现的概率[16]。假设一个句子 S 表示为序列 $S=w_1w_2\cdots w_n$,那么句子 S 出现的概率 $P(S)$ 可以用式(2)表示:

$$p(S)=\prod_{i}^{n}p(w_i\mid w_1w_2\cdots w_{i-1}) \tag{2}$$

由于这个概率的计算量非常庞大,因此将所有历史 $w_1w_2\cdots w_{i-1}$ 按照某个规则映射到等价类 $S(w_1w_2\cdots w_{i-1})$,等价类的数目远远小于不同历史的数目。即假定:

$$p(w_i\mid w_1w_2\cdots w_{i-1})=p(w_i\mid S(w_1w_2\cdots w_{i-1})) \tag{3}$$

根据最大似然估计,式(3)可以表示为:

$$p(w_i\mid w_1w_2\cdots w_{i-1})=\frac{C(w_1w_2\cdots w_{i-1}w_i)}{C(w_1w_2\cdots w_{i-1})} \tag{4}$$

其中,$C(w_1w_2\cdots w_{i-1}w_i)$ 表示 $w_1w_2\cdots w_{i-1}$ 在训练数据中出现的次数。

一般来说,高阶模型能更好地刻画语言的结构,但依旧存在数据稀疏问题。我们采用 Good-Turing 估计法的补偿平滑技术来克服以上问题。根据 Good-Turing 估计法,对于任何出现 r 次的 N-Gram 特征,我们假定它会发生 r^* 次。

$$r^*=(r+1)\frac{n_{r+1}}{n_r} \tag{5}$$

其中,n_r 表示在训练数据中精确出现 r 次 N-Gram 特征的次数。该特征出现的概率为:

$$p(w)=\frac{r^*}{M} \tag{6}$$

其中,M 为计算机病毒特征库中所有特征的总数,定义 $d_r=\frac{r^*}{r}$ 为体贴系数。如果特征 w_1w_2 出现在特征库中,运用 Good-Turing 估计法,条件概率 $p(w_2\mid w_1)$ 可以表示为:

$$p(w_2\mid w_1)=\frac{C^*(w_1w_2)}{C(w_1)}=d_{C(w_1w_2)}\frac{C(w_1w_2)}{C(w_1)} \tag{7}$$

本文将 N-Gram 统计语言模型应用到计算机病毒特征码的二次选择中,其原理是:将病毒程序的一个字节当作一个单词,将计算机病毒候选特征码视为一个句子,可以计算出每个计算机病毒候选特征码的出现概率。出现概率越小,则表示该特征码在其他程序中出现的概率越小,引起的误报率也就越低。因此,通过计算每个候选特征码出现的概率,并将其升序排列,从而选择最小概率的特征码。

4　实验结果与分析

病毒之间通过共性进行分类,主要包括:狭义病毒、木马、蠕虫、后门、Rootkit。通过恶意代码网站[17],对每一类病毒选取 320 个样本;并通过 Windows XP 计算机获得 1600 个合法程序。首先抽取每一类病毒数量的 1/2 样本和合法程序数量的 1/2 样本作为训练集合,用于生成计算机病毒特征库。将其余的样本作为检测样本,提取特征码用以观察本文方法的实验效果。

通过选取不同的 S 值来进行实验,首先取 $S=1000$,然后以 500 为步长,得到的成功率如图 6 所示。

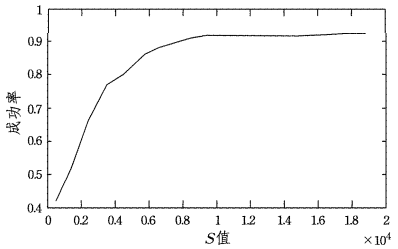

图 6　特征码提取的成功率(SR)

从图 6 可以看出,当 S 取值过小时,成功提取出计算机病毒特征码的几率越小;当 $S\in[9500,18000]$ 时,提取特征码的成功率趋于稳定,说明该范围内的特征基本上来自于病毒程序。对于成功率和误报率的测试,分别使用提取出的计算机病毒特征码对剩余的病毒程序和合法程序进行匹配检测,测试结果如表 2 所列。

表 2　测试的部分结果

病毒名称	病毒特征码	检测出的合法程序数	检测出的病毒程序数	合计
Backdoor. Win32. Ayam	4DA330EE4 5FD2FB1	0	35	35
Worm. Win32. Antinny. ad	93EDC294E A64F021A126	0	28	28
Win32. Ketan	42739051A 05AF515	0	11	11
Win32. Revaz	9B22C0D0E E63C5EAD2	0	27	27
Win32. Htrip. a	E1545173E 1555173	0	39	39
Win32. Alcaul. a	BF36E336E 73 CEB36	0	7	7

从表 2 可以看出,通过该算法提取出的特征码检测出的合法程序数量非常少,但能够检测出一定数目的病毒程序。由此可知,该算法的误报率主要集中在病毒的变种上。

分别使用传统 N-Gram 算法、CVSAE 算法[18]、CSR 算法[19]与本文所提方法对待测样本程序进行特征码提取,以验证本文所提的改进方法的优势。本文主要采用成功率以及误报率作为评价指标,以综合判断特征码提取的性能。判定结果如表 3 所列。

(下转第 361 页)

Security. Magdeburg, Germany, 2004: 160-165.

[15] KITAMURA I, KANAN S, KISHINAMI T. Digital watermarking method for vector map based on wavelet transform [C]// Proceedings of the Geographic Information Systems Association. Tokyo, Japan, 2000: 417-421.

[16] LI Y, XU L. A blind watermarking of vector graphics images [C]// Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications. 2003: 27-30.

[17] 李媛媛,许录平. 矢量图形中基于小波变换的盲水印算法[J]. 光子学报,2004,33(1):97-100.

[18] 易开祥,石教英,孙鑫. 数字水印技术研究进展[J]. 中国图像图形学报,2001,6(2):111-117.

[19] 孙建国. 矢量地图数字水印技术研究[M]. 北京:人民邮电出版社,2012:20-80.

[20] VASSAUX B, NGUYEN P, BAUDRY S, et al. A survey on attacks in image and video watermarking[C]// Proceedings of the SPIE-The International Society for Optical Engineering. 2002: 169-179.

[21] 成科扬. 论集对分析在软件质量评价体系建立中的运用[J]. 计算机应用与软件,2004,21(3):19-21.

(上接第 341 页)

表 3　几种特征码提取方法的分析比较/%

方法	成功率(*SR*)	误报率(*FPR*)
3-Gran	90.76	2.51
4-Gram	91.82	2.73
CVSAE	94.23	2.84
CSR	81.23	3.21
改进的变长 N-Gram	93.21	1.65

由表 3 可知,与其他几种特征码提取方法相比,改进的 N-Gram 特征提取算法可以成功地提取出病毒特征码,并且误报率较低。

结束语　本文采用变长 N-Gram 作为提取特征,同时考虑特征频率的作用,利用特征浓度有向选择生成数据字典,选择有效特征,并将其加入到病毒 N-Gram 特征库中,与待测病毒样本进行匹配;借助 N-Gram 语言模型提取病毒特征码,实现病毒特征码的自动提取工作。从实验结果可以看出,本文提出的基于 N-Gram 改进的病毒特征码提取算法能够有效地提取出计算机病毒特征码,且其性能优于传统的 N-Gram 算法。下一步将研究更为有效的特征选择方法,通过组合使用多个特征码来表示一类病毒,实现病毒的分类检测。

参 考 文 献

[1] YEGNESWARAN V, GIFFIN J T, BARFOD P, et al. An architecture for generating semantics-aware signatures [C]// Conference on Usenix Security Symposium. USENIX Association, 2004: 7-7.

[2] LEE H, KIM W, HONG M. Biologically Inspired Computer Virus Detection System[J]. Lecture Notes in Computer Science, 2004, 3141: 153-165.

[3] KIJEWSKI P. Automated Extraction of Threat Signatures from Network Flows[OL]. http://www.first.org/conference/2006/papers/kijewski-piotr-paper.pdf.

[4] KREIBICH C, ROWCROFT J. Honeycomb: creating intrusion detection signatures using honeypots[J]. Acm Sigcomm Computer Communication Review, 2015, 34(1): 51-56.

[5] 张小康,帅建梅,史林. 基于加权信息增益的恶意代码检测方法[J]. 计算机工程,2010,36(6):149-151.

[6] KEPHART J O, ARNOLD W C. Automatic extraction of computer virus signatures [C]// 4th Virus Bulletin International Conference. 1994.

[7] 张福勇. 基于 n-gram 词频的恶意代码特征提取方法[J]. 网络安全技术与应用,2015(11):88-89.

[8] 白金荣,王俊峰,赵宗渠. 基于 PE 静态结构特征的恶意软件检测方法[J]. 计算机科学,2013,40(1):122-126.

[9] RAFF E, ZAK R, COX R, et al. An investigation of byte n-gram features for malware classification[J]. Journal of Computer Virology & Hacking Techniques, 2016: 1-20.

[10] 曾键,赵辉. 一种基于 N-Gram 的计算机病毒特征码自动提取方法[J]. 计算机安全,2013(10):2-5.

[11] 李沁蕾,王蕊,贾晓启. OSN 中基于分类器和改进 n-gram 模型的跨站脚本检测方法[J]. 计算机应用,2014,34(6):1661-1665.

[12] DHAYA R, POONGODI M. Detecting software vulnerabilies in android using static analysis[C]// International Conference on Advanced Communication, Control and Computing Technologies. 2014.

[13] O'KANE P, SEZER S, MCLAUGHLIN K. N-gram density based malware detection[C]// Computer Applications & Research. IEEE, 2014: 1-6.

[14] SHABTAI A, MOSKOVITCH R, FEHER C, et al. Detecting unknown malicious code by applying classification techniques on OpCode patterns[J]. Security Informatics, 2012, 1(1): 1-22.

[15] SANTOS I, BREZO F, UGARTE-PEDRERO X, et al. Opcode sequences as representation of executables for data-mining-based unknown malware detection[J]. Information Sciences, 2013, 231(9): 64-82.

[16] 吴军. 数学之美[M]. 北京:人民邮电出版社,2012.

[17] 恶意代码网站[OL]. http://vxheaven.org.

[18] 金雄斌. 计算机病毒特征码自动提取技术的研究[D]. 武汉:华中科技大学,2011.

[19] TANG Y, XIAO B, LU X. Using a bioinformatics approach to generate accurate exploit-based signatures for polymorphic worms[J]. Computers & Security, 2009, 28(8): 827-842.