第34卷 第20期 计 算 机 工 程 2008年10月
Vol.34 No.20 Computer Engineering October 2008
·安全技术· 文章编号：1000—3428(2008)20—0159—02 文献标识码：A 中图分类号：TP301.6

基于 Bayes 决策的密码算法识别技术

李继中，蒋烈辉，尹 青，刘铁铭，郭 佳
(解放军信息工程大学信息工程学院，郑州 450002)

摘 要：在可执行代码中识别密码算法对于查找恶意代码、保护计算机系统安全有着重要的意义。该文在对汇编级密码算法特征分析的基础上，提出了汇编级密码算法特征度量元的概念，并采用七维特征向量对其进行形式化描述，建立基于 Bayes 决策的密码算法识别模型，测试结果表明该模型稳定、准确，能够高效地定位代码中的密码算法。

关键词：算法识别；程序理解；决策模型；特征度量元

Cryptogram Algorithm Recognition Technology Based on Bayes Decision-making

LI Ji-zhong, JIANG Lie-hui, YIN Qing, LIU Tie-ming, GUO Jia
(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002)

【Abstract】Recognizing cryptogram algorithm from executable codes plays an important significance in checking malicious codes and protecting computer system. This paper brings forward the concept of assemble cryptogram algorithm characteristic-measurement based on analyzing a lot of assemble cryptogram algorithms, and using seven-dimension characteristic vector to describe it, then constructs a cryptogram algorithm recognition model based on Bayes decision-making. The testing shows that this model is scalable and exact.

【Key words】algorithm recognition; program understanding; decision-making model; haracteristic-measurement

1 概述

在计算机安全防护方面，随着各种杀毒软件的功能日趋强大，恶意代码也曾出现不同的攻击机制。计算机病毒和木马常采用密码算法保护机制隐藏其静态特征，在可执行代码中提取密码算法，对于恶意代码的检测、保护计算机系统的安全性有积极的作用。

算法识别采用的主要技术包括程序依赖图解析、Decomposition Slicing(DS)技术[1]、Abstract Syntax Tree(AST)匹配等。在针对可执行代码的算法识别技术研究方面，主要采用的是特征码匹配技术，该技术对于特征变形、代码混淆效果并不太好。

2 密码算法汇编级统计特征分析

密码算法识别的处理对象二进制目标代码，在此层次上密码算法特征体现并不是很明显，因此借助交互式反汇编工具 IDA 对代码进行处理，在汇编级对密码算法进行特征提取。

2.1 概念和术语

算法的核心处理函数在汇编代码中表现出诸多与一般子程序不同的特征，密码算法特征指标是密码算法核心函数[2-3]与一般子程序有较大区别的各种参数。

度量元是程序的一些基本统计参数，是可以直接度量的特性。基本度量元如表 1 所示。

表 1 子程序基本度量元

编号	度量元名称	描 述
1	子程序的指令条数	子程序的规模
2	子程序的基本块个数	子程序的控制结构
3	逻辑运算类指令条数	子程序的组成
4	算术运算类指令条数	子程序的组成
5	JMP 指令条数	子程序控制转移指令
6	堆栈操作指令条数	包括 push, pop 等指令

根据应用的需要，可以在定义基本度量元的基础上定义新的度量元，以便作为待识别算法的特征指标。

2.2 度量元的提取

选定适合区分密码算法子程序与一般子程序的度量元，是取得良好分类效果的关键。在对大量密码算法子程序统计的基础上，采用以下 4 项指标作为度量元。

(1)汇编子程序的指令条数

通过对密码算法核心处理子程序的指令条数进行统计，发现子程序规模相当大。如MD5 算法子程序的指令条数达到 809 条。选取子程序指令条数为特征指标度量元，记为N_I。

(2)子程序指令使用频度

密码算法子程序用MOV指令完成对S盒的查表操作，用操作类指令完成数据的混淆处理，这 2 类指令在子程序中的使用频度比一般子程序要高出很多。把这 2 类指令的使用频度作为特征指标度量元。设子程序中指令总条数为N_I，MOV指令条数为N_{MOV}，操作类指令条数为N_{OI}，逻辑运算类指令条数为N_{BI}，MOV指令使用频度为R_{MOV}，操作类指令使用频度为R_{OI}，则：$R_{MOV}=N_{MOV}/N_I$, $R_{OI}=N_{OI}/N_I$。

(3)基本块信息

基本块是指一个连续执行的指令序列。密码算法子程序存在主要包含MOV指令和操作类指令的基本块，选取子程序最大基本块的指令条数、基本块内MOV指令和运算类指令使

基金项目：国家“863”计划基金资助项目(2006AA01Z409)
作者简介：李继中(1983－)，男，硕士研究生，主研方向：计算机应用技术，软件逆向工程；蒋烈辉，教授、博士；尹 青，副教授、博士；刘铁铭，讲师；郭 佳，硕士研究生
收稿日期：2007-11-13 **E-mail：**zhongzhong_hero@163.com

用频度作为特征指标度量元。设最大基本块内的指令条数为N_{BI}，MOV指令条数为N_{BMOV}，操作类指令条数为N_{BOI}，MOV指令使用频度为R_{BMOV}，操作类指令使用频度为R_{BOI}，则：$R_{BMOV}=N_{BMOV}/N_{BI}$, $R_{BOI}=N_{BOI}/N_{BI}$。

(4)相同模式代码循环出现

密码算法在进行处理数据过程中，一般使用多轮函数对数据加密处理，轮函数在反汇编结果中的循环模式较为明显。如MD5算法核心函数中，如下模式代码在子程序中循环出现了64次。将子程序是否包含循环模式(*LP*)作为特征指标。

```
0040135A    mov     ecx, ebx
0040135C    not     eax
0040135E    and     eax, [ebp+var_8]
00401361    and     ecx, edi
00401363    mov     edx, edi
00401365    or      eax, ecx
00401367    mov     ecx, [ebp+arg_0]
0040136A    add     eax, [ebp+var_48]
0040136D    lea     ecx, [ecx+eax-28955B88h]
00401374    rol     ecx,7
00401377    mov     eax, ecx
00401379    add     eax, edi
```

综合上面4个汇编级密码算法度量元，针对密码算法子程序的特点，定义一个七维特征向量$\boldsymbol{EV}(N_I, R_{MOV}, R_{OI}, N_{BI}, R_{BMOV}, R_{BOI}, LP)$，该特征向量是进行密码算法识别的前提。

3 Bayes密码算法识别模型

Bayes分类模型[4]是一种简单、有效的分类器，该模型具有逻辑简单、算法实施的时间空间开销小等优点。

3.1 Bayes分类模型

样本有n个属性$A_1,A_2,\cdots,A_n$，每个样本可看作是n维空间的一个点$X=(x_1,x_2,\cdots,x_n)$。其中，$x_1,x_2,\cdots,x_n$分别为样本属性$A_1,A_2,\cdots,A_n$的取值($n\in\mathbb{Z}$)。

有m个不同的类别$C_1,C_2,\cdots,C_m$($m\in\mathbb{Z}$)，X是一个未知类别的样本，$P(C_i)$表示样本属于类别C_i的概率，$P(C_i\mid X)$和$P(C_j\mid X)$则分别表示样本的属性取值为X的条件下，样本属于类别C_i和C_j的概率。Bayes分类模型将未知类别的样本X归属到类别C_i，当且仅当$P(C_i\mid X)>P(C_j\mid X)$对于所有的$j$成立($j\neq i, i\le m, j\le m, i,j\in\mathbb{Z}$)。

设S_i表示训练样本中属于类别C_i的样本个数，S表示全部训练样本的样本个数。由Bayes定理可得：$P(C_i\mid X)=P(X\mid C_i)P(C_i)/P(X)$，其中，$P(C_i)=S_i/S$。设各类别相互独立，则：$P(X\mid C_i)=P(x_1\mid C_i)P(x_2\mid C_i)\cdots P(x_n\mid C_i)$。

3.2 Bayes分类模型在密码算法识别中的应用

设有m个密码算法子程序(CP)和n个一般子程序(NP)($m,n\in\mathbb{Z}$)。确定k个特征指标($k\in\mathbb{Z}$)。将m个CP中满足第i个特征指标($k\ge i>0, i\in\mathbb{Z}$)的CP个数记为C_{CPi}。

为了对实时数据进行分类，定义密码算法的特征阈值，以此度量子程序是否满足该项特征指标。当统计结果分布较为分散时，取平滑数$N_\alpha=(\sqrt[2]{N_1^2+N_2^2+\cdots+N_n^2})/n$作为阈值；当统计结果分布较为集中时，取几何平均数$N_\alpha=\sqrt[n]{N_1\cdot N_2\cdots\cdots N_n}$作为阈值。

设T表示某个子程序是CP的事件，SC_i表示该子程序满足第i个特征指标的事件。则：

$$P(SC_i\mid T)=C_{CPi}/m,\ P(\overline{SC_i}\mid T)=1-P(SC_i\mid T)$$

设F表示子程序是一般子程序的事件。将n个一般子程序中满足第i个特征指标的子程序个数记为C_{NPi}。在子程序是NP的情况下，将满足第i个特征指标C_i事件的概率记为$P(SC_i\mid F)$，则：$P(SC_i\mid F)=C_{NPi}/n$, $P(\overline{SC_i}\mid F)=1-P(SC_i\mid F)$。

样本中某个子程序是CP的概率为$P(T)=m/(m+n)$，是NP的概率为$P(T)=n/(m+n)$。

现有一个实时子程序，该子程序满足了已选定的k个特征指标中的i个特征指标($0<i\le k, i\in\mathbb{Z}$)，不满足另外的(k-i)个特征指标，须判定其是否为CP。

设X_j ($0<j\le k, j\in\mathbb{Z}$)表示满足第j个特征指标的事件，子程序样本可表示为k维空间的一个点$X=(X_1,\cdots,X_i,\cdots,\overline{X_j},\cdots,X_k)$，设样本在同时满足$i$个特征指标且不满足另($k$-$i$)个特征指标的情况下是CP的概率为$P(T\mid X)$，是NP的概率为$P(F\mid X)$。由Bayes定理得：

$$P(T\mid X)=P(X\mid T)\cdot P(T)/P(X)$$
$$P(F\mid X)=P(X\mid F)\cdot P(F)/P(X)$$

3.3 数据训练

根据选取的特征指标，对密码算法子程序和一般子程序进行了数据统计，根据统计数据确定N_I, R_{MOV}, R_{OI}, N_{BI}, R_{BMOV}和R_{BOI}的阈值。

根据密码算法子程序的统计结果，N_I和N_{BI}特征指标取统计结果的平滑数作为阈值：

$$N_{I\alpha}=(\sqrt[2]{N_1^2+N_2^2+\cdots+N_n^2})/n\approx 210,\ N_{BI\alpha}\approx 200$$

R_{MOV}, R_{OI}, R_{BMOV}, R_{BOI}特征指标取统计结果的几何平均数作为阈值：

$$R_{MOV\alpha}=\sqrt[n]{N_1\cdot N_2\cdots\cdots N_n}\approx 0.375\ 4,$$
$$R_{OI\alpha}\approx 0.456\ 1,\quad R_{BMOV\alpha}\approx 0.450\ 6,\quad R_{BOI\alpha}\approx 0.429\ 4$$

根据训练数据，计算$P(SC_i\mid T)$，$P(SC_i\mid F)$, $P(T)$和$P(F)$，其中，$P(SC_i\mid T)$, $P(SC_i\mid F)$ ($1\le i\le 7, i\in Z$)分别和$N_{I\alpha}$, $R_{MOV\alpha}$, $R_{OI\alpha}$, $R_{BI\alpha}$, $R_{BMOV\alpha}$, $R_{BOI\alpha}$, LP相对应。例如在训练子程序中，共有3个满足指令条数大于阈值，则$P(SC_1\mid F)=0.15$；同理可得：$P(SC_1\mid T), P(SC_1\mid F), \cdots, P(SC_7\mid T), P(SC_7\mid F)$，先验概率$P(T)\approx 0.285\ 7$, $P(F)\approx 0.714\ 3$。

4 测试

为验证Bayes密码算法识别模型的正确性和可靠性，对常用的应用程序进行筛选分析，测试平台为：Windows 2000操作系统，Pentium4处理器，512 MB内存。测试结果如表2所示。

表2 密码算法核心函数筛选测试结果

扫描软件名称	文件大小/B	子程序个数	特征收集时间/s	筛选时间/s	密码算法核心函数筛选定位结果
MD5calculator.exe	32 768	64	1.281	1.250	SUB_40132A
keygen.exe	36 864	74	1.297	1.266	SUB_4012B0
bfkeygen.exe	69 632	85	1.921	1.214	SUB_401130
msconfig.exe	151 552	294	6.023	1.326	无
Unacev2.dll	75 264	457	9.024	1.057	SUB_40B1F2 SUB_40B49D

结果表明分类程序能够迅速定位反汇编结果文件中所包含的密码算法核心函数。在测试过程中，虽然特征采集时间会随着目标代码的增加而线形增加，但得益于Bayes分类模型简单有效的特点，筛选时间较短。

(下转第163页)