计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190393
56(11):2339-2348,2019

一种基于概率主题模型的恶意代码特征提取方法

刘亚姝[1,2] 王志海[1] 侯跃然[3] 严寒冰[4]
[1](北京交通大学计算机与信息技术学院 北京 100044)
[2](北京建筑大学电气与信息工程学院 北京 100044)
[3](北京邮电大学网络技术研究院 北京 100876)
[4](国家计算机网络应急技术处理协调中心 北京 100029)
(ly_s8020@163.com)

A Method of Extracting Malware Features Based on Probabilistic Topic Model

Liu Yashu[1,2], Wang Zhihai[1], Hou Yueran[3], and Yan Hanbing[4]
[1](*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)
[2](*School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing* 100044)
[3](*Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing* 100876)
[4](*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029)

Abstract In the current complex network environment, malicious codes have been spread quickly in various ways, which illegally occupy user terminal equipment or network equipment and illegally steal privacy data. Malware poses a serious security threat to network and Internet users. Traditional methods can't detect unknown malicious codes which is challenged by the diversity and large number of malicious code variants. We propose an unsupervised malware identification approach that generates a standardization rule of assembly instructions by analyzing the content of the decompiled PE files. By introducing latent Dirichlet allocation (LDA), our method extracts the latent "document-topic" and "topic-word" probability allocation from samples. The topic probability distributions are extracted as features of samples, which is a new way for malware feature presentation. Then, we propose a new malware detecting framework to train model and test malware. What's more, our method solves the problem that the topic number in LDA model needs to be specified beforehand using the perplexity and different steps, which evaluates the best numbers of "topics" quickly and automatically. Finally, it analyzes the semantics of "document-topic" and "topic-word" aggregating results in assembly instructions, which explains the latent semantics of features obtained by our method. Experimental results show that our method is more discriminative, which has better classification results than other methods, while providing accurate discrimination of the new novel malware variants.

Key words malware detection; latent Dirichlet allocation (LDA); probabilistic topic model; perplexity; Gibbs

摘 要 在当前复杂网络环境下,恶意代码通过各种方式快速传播,入侵用户终端设备或网络设备、非法窃取用户隐私数据,对网络和互联网用户造成了严重的安全威胁.传统检测方法难以检测未知恶意

收稿日期:2019-06-11;**修回日期**:2019-08-06
基金项目:国家重点研发计划项目(2018YFB0803604,2018YFB0804704);国家自然科学基金项目(U1736218,61672086)
This work was supported by the National Key Research and Development Program of China (2018YFB0803604, 2018YFB0804704) and the National Natural Science Foundation of China (U1736218, 61672086).

代码,而恶意代码变体的多样性和庞大数量也对未知恶意代码检测构成了巨大挑战.提出了一种无监督的恶意代码识别方法,通过分析反汇编 PE 文件给出汇编指令标准化规则,结合潜在狄立克雷分布(latent Dirichlet allocation, LDA)获得汇编指令中潜在的"文档-主题"、"主题-词"的分布.再以"主题分布"构造恶意样本特征,产生一个全新的恶意代码检测框架.结合"困惑度"和变化的步长给出了最优"主题"数目的快速评价和自动确定方法,解决了 LDA 模型中主题数目需要预先指定的问题.同时解析了"文档-主题"、"主题-词"聚集结果的语义可解释性,说明了该方法获得的样本特征具有潜在的语义.实验结果表明:与其他方法相比该方法具有相当的或更好的恶意代码鉴别能力,同时能够准确地识别恶意代码的新变体.

关键词 恶意代码检测;狄立克雷分布;概率主题模型;困惑度;Gibbs

中图法分类号 TP393

恶意代码是各种类型恶意软件的统称,包括病毒、特洛伊木马、后门、蠕虫等.据 2019 年 Symantec 发布的《互联网安全威胁报告》称,全球日均拦截威胁数量达 142 亿个,几乎每一种物联网设备都很容易遭受攻击[1].恶意代码对互联网企业、个人用户的数据安全和财产安全造成了极大的威胁.

最常见的恶意代码检测方法是基于特征码的检测,通过人工提取、构造特征库,比对相同位置的字节码来判断样本是否为恶意代码[2].这种方法被各大反病毒工具广泛使用,通过不断更新特征库以提高保护企业、个人用户信息安全的能力.但是基于特征码的检测方法不能识别未知的恶意代码,而随着各种开发工具的发展,恶意代码的变体越来越多样、反检测能力越来越强,使得各大反病毒和安全厂商面临着巨大的挑战.在与恶意代码博弈中,研究人员也提出很多有价值的研究成果.

Moskovitch 等人针对反汇编后的文件指令、结构等从语法和语义角度分析汇编代码,提出以 n-grams 操作码序列为特征构造恶意代码的特征集[3];Santos 等人分析操作码的频度以达到检测未知恶意样本的目的[4].Kapoor 将操作码与控制流程图(control flow graph, CFG)结合起来实现多类别的分类[5].Ashkan 提出基于 API 调用的检测方法[6].

2001 年 Matthew 将数据挖掘技术引入恶意代码检测中,恶意代码检测技术有了飞速地发展,传统恶意代码检测技术与数据挖掘技术相结合产生了更好的检测效果[7-8].Saxe 等人引入深度神经网络可以实现大规模恶意样本的低误报率检测[9].以 Nataraj 为代表的研究人员另辟蹊径,将二进制可执行 PE 文件转化为灰度图像,借助图像处理的办法和机器学习算法实现恶意样本的分类问题[10-13].

Tamersoy 等人考虑样本之间的关系图[14];Ye 等人结合样本内容和样本之间的关系实现云端恶意代码的检测[15];Fan 等人将样本与 API、样本与压缩包、样本与机器之间关系构造为异构信息网络以检测应用程序的恶意性[16].

由于恶意代码相对于良性代码而言会有特殊的行为,例如会有特定的访问序列、特定的行为以及对内存的控制,因此基于行为的分析是动态检测中常用的技术[17-20].

综上,无论是动态还是静态分析方法,通过分析样本的内容、样本之间的关系提取恶意特征,并采用机器学习的方法分类恶意样本是一种常用的方法.

本文基于 Windows 平台,从静态分析入手针对恶意代码样本的反汇编文本,采用概率主题模型聚集样本特征,并用机器学习算法实现恶意样本的分类问题.本文主要贡献有 3 个方面:

1) 给出了汇编指令标准化规则.该规则不仅针对操作码,同时也考虑了操作数的影响.细化后的规则可以提高 0.4 左右的分类精确率.

2) 提出了具有潜在语义的特征表示方法.本文引入概率主题模型——LDA,通过计算潜在的"文档-主题"概率分布得到样本特征.该特征具有潜在语义信息,更具有鉴别能力.

3) 给出了一种无监督学习模型.本文结合 LDA 模型提出了一种全新的、无监督的学习模式,它可以为新样本赋予与训练集相关的主题概率分布,具有检测新样本的能力.同时将困惑度与不定步长相结合快速、准确选取合适的"主题"数目,解决 LDA 模型需要预先设定主题数目的问题.

本文为恶意代码的检测提出了一个新的研究视角.实验表明:与其他方法相比,本文方法不仅具有较好的分类性能,而且能够识别新的恶意代码.

1 概率主题模型

概率主题模型是一种统计方法,以非监督学习的方式分析文本中的“词”从而发现蕴藏于其中的“主题-词”、“主题-文档”之间的结构.Christos 等人于 1998 年提出了潜在语义索引(latent semantic indexing, LSI)[21].Hofmann 于 1999 年提出了概率潜在语义分析(probabilistic latent semantic analysis, PLSA)[22].随后,Blei 等人于 2003 年提出了潜在狄利克雷分布(latent Dirichlet allocation, LDA)[23].LDA 是一个生成模型,也是一个 3 层贝叶斯概率模型,包含词、主题和文档 3 层结构,可以生成“文档-主题”模型.在主题模型中,“主题(topic)”就是文本中“词(word)”的条件概率分布,表示“词”与“主题”之间的关联性,反映了“词”在该“主题”下出现的频繁程度,即与该“主题”关联性高的“词”有更大概率出现.其概率模型图如图 1 所示:

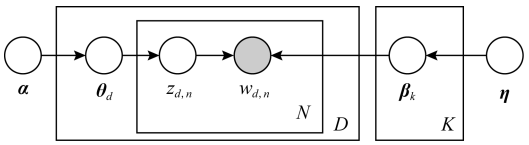


Fig. 1 Probabilistic model diagram of LDA

图 1 LDA 概率模型图

1.1 模型定义

在主题模型中,涉及到“文档”、“主题”、“词”,下面给出符号定义.一篇“文档”W 是由 N 个“词”w_i 构成,即 $W=\{w_1,w_2,\cdots,w_N\}$ 构成,这里 w_i 表示第 i 个“词”;一个“语料库”D 是 M 篇文档的集合,即 $D=\{d_1,d_2,\cdots,d_M\}$,语料库中包含的全部“词”可以表示为 $W=\{w_1,w_2,\cdots,w_V\}$;“主题”是隐含在文档中,用于聚集相关的“词”,可以表示为 $Z=\{z_1,z_2,\cdots,z_K\}$.概率主题模型,不仅能够对语料库中的“文档”而且对于其他“相似文档”都具赋予较高的概率.LDA 模型中“主题”分布与“词”分布都服 Dirichlet 先验分布.

根据图 1,产生一篇文档的方式为:

1) 采样主题 d 的主题分布 $\boldsymbol{\theta}_d\sim Dirichlet(\boldsymbol{\alpha})$,得到文档 d 的主题分布;

2) 生成第 d 个文档的第 n 个主题 $z_{d,n}\sim Multinomail(\boldsymbol{\theta}_d)$,得到词的主题;

3) 采样主题 $z_{d,n}$ 的词的分布 $\boldsymbol{\beta}_k\sim Dirichlet(\boldsymbol{\eta})$,得到该主题的词分布;

4) 采样生成文档 d 的第 n 个主题下的词 $w_{d,n}\sim Multinomail(\boldsymbol{\beta}_{z_{d,n}})$.

这里,$\boldsymbol{\alpha}$ 为 Dirichlet 分布的 K 维超参,用于产生任一文档 d_i 的主题分布 $\boldsymbol{\theta}_i$;$\boldsymbol{\eta}$ 为 Dirichlet 分布的 V 维超参,V 代表 D 中“词”的个数.本文中使用的符号及说明如表 1 所示:

Table 1 Notations in this Paper

表 1 本文使用符号说明表

Notation	Meaning
K	The number of topics
M	The number of documents
V	The number of words
N_d	The number of words in the dth document
$\boldsymbol{\theta}_d$	Multinomial distribution of “document-topic” in the dth document, $\boldsymbol{\Theta}=(\boldsymbol{\theta}_d)_{d=1}^{M}$ is $M\times K$ matrix
$\boldsymbol{\beta}_k$	Multinomial distribution of “topic-word”, $\boldsymbol{\Phi}=(\boldsymbol{\beta}_k)_{k=1}^{K}$ is $K\times V$ matrix
$z_{d,n}$	The nth topic of the dth document
$w_{d,n}$	The nth word of the dth document
$\boldsymbol{\alpha}$	The super parameter of Dirichlet distribution, K-dimensional vector
$\boldsymbol{\eta}$	The super parameter of Dirichlet distribution, V-dimensional vector Dirichlet

LDA 模型的目标是找到每一篇文档的“主题分布”和每一个主题中“词的分布”.在模型中,需要预先假定主题数目——“K”.LDA 模型中,假设文档中的“词”是无序的、互相独立的.文档中的词,通过统计词频构成词向量,也即“词包(bag of words)”.“主题”产生“词”不依赖具体某一个文档,因此“文档-主题”分布和“主题-词”分布是相互独立的.

1.2 模型推导

为了获得“文档-主题”和“主题-词”的概率分布,也即 $\boldsymbol{Z},\boldsymbol{W}$ 的概率分布,可以采用 Gibbs 采样的方法.

由于 $P(\boldsymbol{W},\boldsymbol{Z})\propto P(\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{\alpha},\boldsymbol{\eta})$,根据 Dirichlet 分布与多项式分布共轭的特性,可以简化条件概率 $P(\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{\alpha},\boldsymbol{\eta})$ 的求解:

$$P(\boldsymbol{W},\boldsymbol{Z}\mid\boldsymbol{\alpha},\boldsymbol{\eta})=P(\boldsymbol{W}\mid\boldsymbol{Z},\boldsymbol{\eta})\ P\ (\boldsymbol{Z}\mid\boldsymbol{\alpha})=\int P(\boldsymbol{W}\mid\boldsymbol{\Phi},\boldsymbol{Z})\ P\ (\boldsymbol{\Phi}\mid\boldsymbol{\eta})\ \mathrm{d}\boldsymbol{\Phi}\times\int P(\boldsymbol{Z}\mid\boldsymbol{\Theta})\ P\ (\boldsymbol{\Theta}\mid\boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\Theta},\tag{1}$$

有:

$$\int P(\boldsymbol{Z}\mid\boldsymbol{\theta})\ P\ (\boldsymbol{\theta}\mid\boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\theta}=\int\prod_{d=1}^{M}\Big(\prod_{k=1}^{K}\theta_{k,d}^{n_d^{(k)}}\ \mathrm{Dirichlet}(\boldsymbol{\theta}\mid\boldsymbol{\alpha})\ \mathrm{d}\boldsymbol{\theta},\tag{2}$$

这里,$n_d^{(k)}$ 表示在第 d 个文档中第 k 个主题的词的个数表示,进一步可得:

$$\int P(\boldsymbol{Z}\mid\boldsymbol{\Theta})\,P(\boldsymbol{\Theta}\mid\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\Theta}=\int\prod_{d=1}^{M}\Delta(\boldsymbol{\alpha})\prod_{k=1}^{K}\theta_{k,d}^{n_d^{(k)}+\alpha_k-1}\mathrm{d}\boldsymbol{\Theta},\tag{3}$$

这里,$\Delta(\boldsymbol{\alpha})=\dfrac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}$.

由于 Dirichlet 分布与多项式分布共轭则式(3)可写为

$$\int P(\boldsymbol{Z}\mid\boldsymbol{\Theta})\,P(\boldsymbol{\Theta}\mid\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\Theta}=\prod_{d=1}^{M}\Delta(\boldsymbol{\alpha})\frac{\prod_{k=1}^{K}(n_d^{(k)}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(n_d^{(k)}+\alpha_k)\right)},\tag{4}$$

同理可得,式(1)中,

$$\int P(\boldsymbol{W}\mid\boldsymbol{\Phi})\,P(\boldsymbol{\Phi}\mid\boldsymbol{\eta})\,\mathrm{d}\boldsymbol{\Phi}=\prod_{k=1}^{K}\Delta(\boldsymbol{\alpha})\frac{\prod_{v=1}^{V}(n_k^{(v)}+\eta_v)}{\Gamma\left(\sum_{v=1}^{V}(n_k^{(v)}+\eta_v)\right)},\tag{5}$$

这里 $n_k^{(v)}$ 表示第 k 个主题中第 v 个词的个数,则有:

$$P(\boldsymbol{W},\boldsymbol{Z}\mid\boldsymbol{\alpha},\boldsymbol{\eta})=\prod_{k=1}^{K}\Delta(\boldsymbol{\eta})\frac{\prod_{v=1}^{V}(n_k^{(v)}+\eta_v)}{\Gamma\left(\sum_{v=1}^{V}(n_k^{(v)}+\eta_v)\right)}\times\prod_{d=1}^{M}\Delta(\boldsymbol{\alpha})\frac{\prod_{k=1}^{K}(n_d^{(k)}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(n_d^{(k)}+\alpha_k)\right)}=\left(\frac{\Gamma\left(\sum_{v=1}^{V}\eta_v\right)}{\prod_{v=1}^{V}\eta_v}\right)^{K}\times\left(\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\alpha_k}\right)^{M}\times\prod_{k=1}^{K}\frac{\prod_{v=1}^{V}(n_k^{(v)}+\eta_v)}{\Gamma\left(\sum_{v=1}^{V}(n_k^{(v)}+\eta_v)\right)}\times\prod_{d=1}^{M}\frac{\prod_{k=1}^{K}(n_d^{(k)}+\alpha_k)}{\Gamma\left(\sum_{k=1}^{K}(n_d^{(k)}+\alpha_k)\right)}.\tag{6}$$

这样,获得了 $P(\boldsymbol{W},\boldsymbol{Z})$ 联合概率分布的近似形式.在 $P(\boldsymbol{W},\boldsymbol{Z})$ 中词向量 $\boldsymbol{W}$ 是已知的,因此:

$$P(z_{d,n}\mid\boldsymbol{W},\boldsymbol{Z}_{\neg d,n},\boldsymbol{\alpha},\boldsymbol{\eta})=\frac{P(\boldsymbol{W},\boldsymbol{Z}\mid\boldsymbol{\alpha},\boldsymbol{\eta})}{P(\boldsymbol{W}_{\neg d,n},\boldsymbol{Z}_{\neg d,n}\mid\boldsymbol{\alpha},\boldsymbol{\eta})\,P(w_{d,n}\mid\boldsymbol{\alpha},\boldsymbol{\eta})}\propto\frac{n_{z_{d,n}}^{(w_{d,n})}+\eta_{w_{d,n}}-1}{\sum_{v=1}^{V}(n_{z_{d,n}}^{(v)}+\eta_v-1)}\times(n_d^{(z_{d,n})}+\alpha_{d,n}-1),\tag{7}$$

其中,$\boldsymbol{Z}_{\neg d,n}$ 表示去掉词 $w_{d,n}$ 的主词分布,$\boldsymbol{W}_{\neg d,n}$ 表示词向量 $\boldsymbol{W}$ 中去掉词 $w_{d,n}$,由于:

$$P(\boldsymbol{\theta}_d\mid\boldsymbol{Z}_d,\boldsymbol{\alpha})=\mathrm{Dirichlet}(\boldsymbol{\theta}_d\mid\boldsymbol{\alpha}+n_d),\tag{8}$$

$$P(\boldsymbol{\beta}_k\mid\boldsymbol{Z},\boldsymbol{W},\boldsymbol{\eta})=\mathrm{Dirichlet}(\boldsymbol{\beta}_k\mid\boldsymbol{\eta}+n_k).\tag{9}$$

根据 Dirichlet 分布的期望公式,对式(8)(9)取数学期望,可以得到:

$$\beta_{k,v}=\frac{n_k^{(v)}+\eta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\eta_v)},\tag{10}$$

$$\theta_{d,k}=\frac{n_d^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_d^{(k)}+\alpha_k)}.\tag{11}$$

Gibbs 采样"词"的"主题",即得到所有词的采样主题,进而得到 $\boldsymbol{\theta}_d$,$\boldsymbol{\beta}_k$.

2 恶意样本特征的提取

一个恶意样本可以通过逆向工具将其 PE 文件转化为汇编语言代码.本文通过 LDA 模型对恶意样本集中的汇编操作指令进行分析、提取其特征,并解决恶意样本的分类问题.

2.1 汇编指令的预处理

本文采用了 python 包中的 pefile 完成对 PE 文件的解析,获得汇编文件,如图 2 所示:

```
.text:01027940          mov     edi, edi
.text:01027942          push    ebp
.text:01027943          mov     ebp, esp
.text:01027945          push    ecx
.text:01027946          push    ebx
.text:01027947          xor     ebx, ebx
.text:01027949          cmp     dword ptr [ebp+ArgList], ebx
.text:0102794C          mov     [ebp+hKey], ebx
.text:0102794F          jz      loc_1027A26
.text:01027955          cmp     [ebp+Str], ebx
.text:01027958          jz      loc_1027A26
.text:0102795E          push    esi
.text:0102795F          lea     eax, [ebp+hKey]
.text:01027962          push    eax             ; phkResult
.text:01027963          push    3               ; samDesired
.text:01027965          push    ebx             ; ulOptions
.text:01027966          mov     esi, offset aSoftwareMicr_3 ;
.text:0102796B          push    esi             ; lpSubKey
.text:0102796C          push    80000002h       ; hKey
.text:01027971          call    sub_101C02D
.text:01027976          cmp     eax, ebx
.text:01027978          jz      short loc_1027992
.text:0102797A          lea     eax, [ebp+hKey]
.text:0102797D          push    eax             ; phkResult
.text:0102797E          push    3               ; samDesired
.text:01027980          push    ebx             ; ulOptions
.text:01027981          push    esi             ; lpSubKey
.text:01027982          push    80000001h       ; hKey
.text:01027987          call    sub_101C02D
.text:0102798C          mov     esi, eax
.text:0102798E          cmp     esi, ebx
.text:01027990          jnz     short loc_10279E6
```

Fig. 2 Disassemble example of PE file

图 2 PE 文件反汇编示例

如图 2 所示,经过反汇编解析得到的汇编文件中包含了很多冗余、干扰信息.在此仅针对其中的汇编指令提取特征.一条汇编指令由操作数和操作码构成,本文不仅针对操作码,同时操作数也参与特征的聚集.由于汇编指令的格式复杂、长短差距悬殊、含义高度丰富和杂乱、数量庞大,不能直接作为语料库.为了更好地控制"词典"的数据量,对数据进行粗糙化处理——汇编代码标准化,使其表示的类型有限、表示的规律性更明显.

具体方法为:

1) 操作码对齐

在汇编指令中操作码的长度为 2~6 个字符不等,由于 3 个字符长度以内的操作码占操作码类型的 48.2%、4 个字符长度的操作码占 28.8%、5 个字符长度的操作码占 17.8%、6 个字符长度的操作码占 5.2%.而如果考虑到使用频率,3 个字符长度内的操作码被调用的占比超过 90%.因此,本文选用 3 个字符长度描述操作码.例如 push→pus,mov→mov,call→cal,je→je_(_表示空格)等.

2) 操作数标准化

① 寄存器.由于寄存器的种类较多,常用的有 8 b,16 b 以及 32 b 三种主要的寄存器.如 eax,ebx,ecx,edx,esi,edi,ebp,esp 等寄存器标准化为 r32;ax 标准化为 r16;al 标准化为 rg8[24].

② 内存.标准化为 MEM.如[*eax*],[*edi*+4]等均表示为 MEM.

③ 立即数.标准化为 VAL,如 0,5A4Dh 表示为 VAL.

④ 调用指令.调用外部的系统库函数时指令不做处理;调用内部函数如"call sub_101C02D"时规范化为"call sub".

⑤ 跳转指令后的操作数.如"jz short loc_4023E7"规范化"jz loc".

表 2 给出了按照如上规则标准化前后的代码块的对照.

Table 2 Example of Standardizing the Assembly Instructions

表 2 汇编指令标准化示例

Source Code Block	Standard Code Block
push 70h	pus VAL
xor ebx,ebx	xor r32,r32
push ebx	pus r32
cmp eax,ebx	cmp r32,r32
mov ecx,[eax+3Ch]	mov r32,MEM
add ecx,eax	add r32,r32
jnz short loc_1018887	jnz loc

将每条汇编指令设定为一个"词".图 3 所示的是 MD5 为 0C1BF77A51B6308D62F0743C3B1A9FF1.3AF3EF67 的样本,经过汇编指令规则化后、提取"词"的部分结果.

pus oth,mov ebp oth,sub esp oth,cmp mem oth,je oth,cmp mem oth,mov eax mem,jl oth,cmp mem oth,je oth,pus mem,cal mem,mov mem oth,tes eax oth,je oth,mov edi,inc oth,add eax oth,cmp mem oth,jbe oth,lea edx mem,pus oth,move eax mem,tes rg8 oth,pop oth,pop oth,lea oth,ret oth,move eax mem,tes eax oth,jns oth,inc oth,pus oth,and ecx oth,pus mem,add eax oth,je oth,mov esi oth,jmp oth,mov esi mem,add esi oth,pus oth,cal mem,jmp oth,pus oth,cal oth,lea esi mem,pus oth,pus oth,cal oth,je oth,pus oth,pus oth,cal oth,jmp oth,pus oth,cal oth,mov esi oth,pus oth,cal oth,cmp eax oth,je oth add eax oth,pus oth,pus mem,cal oth,jmp oth,xor eax oth,mov mem oth,je oth,pus oth,pus oth,cal mem,jmp oth,mov eax mem,add eax oth,pus oth,pus oth,xor ecx oth,pus oth,cmp mem oth,jne oth,cal mem,jmp oth,……

Fig. 3 Example of segmentation result in the malicious sample

图 3 恶意样本"分词"结果示例

2.2 检测框架

将训练集中恶意样本按照上述方法标准化后,可统计出训练集的词典、每个样本的"词袋"作为 LDA 模型训练的输入数据,进而得到恶意样本的特征并实现分类.

本文在使用 LDA 模型提取恶意样本特征时,训练集样本会被使用 2 次,第 1 次是用来构建 LDA 模型,第 2 次是用来产生训练集的特征数据.但是,不论训练样本还是测试样本,都需要经过样本预处理.具体的工作流程如图 4 所示.工作流程可以分为 3 个阶段:

1) 训练样本预处理阶段.该阶段产生可被 LDA 模型处理的数据.首先需要标准化训练集样本的汇编指令、提取"词典",计算每个样本的"词包".初始

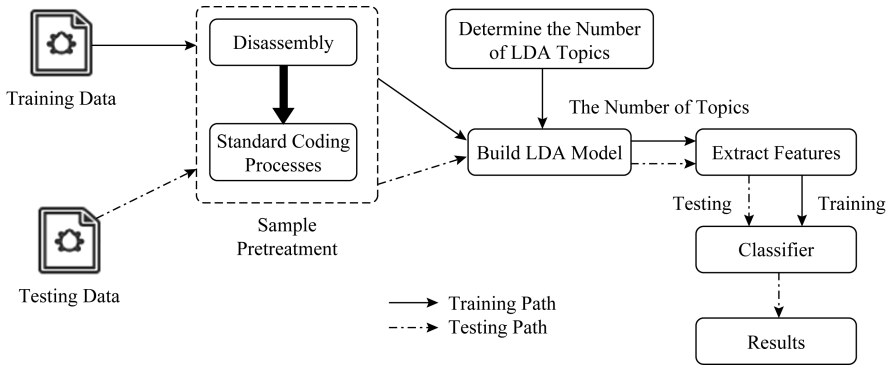


Fig. 4 Workflow diagram of classifying malwares using LDA model

图 4 应用 LDA 模型的恶意代码分类工作流程图

设置 LDA 模型主题个数、输入 LDA 模型其他参数,产生当前数据分布下的 LDA 模型.经过多次困惑度和不定步长评价、选择最优的主题数目,进而得到最优主题数目下 LDA 的训练模型.

2) LDA 建模阶段.样本预处理结果输入 LDA 模型会产生与 LDA 主题数相同维度的特征,每个维度标识在相应主题上的拟合程度,从而得到训练样本的特征,并训练分类器.

3) 分类阶段.该阶段将测试样本经过样本预处理、获得测试集的特征,将特征输入训练好的分类器,得到分类结果.

3 实验与分析

本节中我们构建了基于 LDA 模型的恶意代码检测模型,并在 2 个数据集测试了本文方法.首先选取 2015 年微软 Kaggle 数据集[25],包括训练集、测试集和训练集的标注.其中每个恶意代码样本(去除了 PE 头)包含 2 个文件:一个是十六进制表示的".bytes"文件,另一个是利用 IDA 反汇编工具生成的".asm"文件.

我们首先进行了小样本验证实验.在微软 Kaggle 数据集中,随机选取某一家族为测试集(例如随机选择编号为 7 的家族中的 70 个样本);训练数据为标号为 7 的家族中抽取的 80 个样本,以及其余家族中随机抽取的 80 个样本,共 160 个样本.经实验,当主题数为 5 时采用随机森林(random forest, RF)分类器(参数为 30)精确率为 0.969.证明本文方法对恶意代码的分类是有效的.

第 2 个数据集来自于 CNCERT,包含 10 个家族、15 000 个恶意代码样本.本文在 CNCERT 提供的数据集上测试了我们的方法并完成与他人方法的实验对比.

3.1 汇编指令标准化粗糙程度对分类结果的影响

按照 2.1 节汇编指令标准化规则,在 CNCERT 数据集上,当主题数目为 240 个时本文方法平均分类精确率达到最好,为 0.90.每个家族具体的分类评价指标如表 3 所示:

Table 3 Classification Results of Every Family

表 3 各家族分类结果

Family	Accuracy	Recall	$F1$-score
Agent	0.9	0.71	0.79
Softpulse	0.72	0.99	0.83
Allaple	0.97	1	0.98
Mentiger	0.98	0.99	0.99
Adload	0.91	0.86	0.88
Nimnul	0.90	0.86	0.88
LMN	1	0.99	1
Virut	0.98	0.89	0.93
WBNA	0.94	0.94	0.94
Ageneric	0.80	0.66	0.72
Average	0.90	0.89	0.89

从表 3 可以看出 LMN 家族分类结果最好,分类精确率可以达到 100%;Softpulse 家族和 Ageneric 家族的分类精确率较低.经过对比这 2 个家族的标准化操作码的结果发现,32 b 寄存器均被标准化为 r32,这种标准化方法过于粗糙,不利于分析较敏感的数据.

为此,细化了汇编指令的标准化规则,以便尽可能地抽取其中的信息,同时保证词典不会过于庞大.

寄存器中 32 b 寄存器更多地负责与程序有关的信息,将 32 b 寄存器具体标示出来.这些寄存器有 EAX,EBX,ECX,EDX,ESI/EDI 以及 EBP 等.

此外还有 R0～R3 四个与程序运行息息相关的寄存器,也将它们具体进行标示.但是,这些寄存器并不是所有的编译环境都支持的,在样本中几乎不出现.

细化了寄存器类别后,采用 RF 分类器(参数为 30)、主题数目为 240 时,平均分类精确率为 0.94,相比细化前(0.90)有了较大提高,如表 4 所示:

Table 4 Classification Results After More Specified Register Classes

表 4 细化后各家族分类结果

Family	Accuracy	Recall	*F*1-score
Agent	0.93	0.82	0.87
Softpulse	0.94	0.99	0.96
Allaple	0.98	1	0.99
Mentiger	0.96	1	0.98
Adload	0.95	0.97	0.96
Nimnul	0.88	0.96	0.92
LMN	0.99	1	0.99
Virut	0.97	0.93	0.95
WBNA	0.96	0.95	0.96
Ageneric	0.88	0.81	0.84
Average	0.94	0.94	0.94

从表 4 相比表 3 可以看到,Agent 家族、Softpulse 家族、Ageneric 家族的精确率、召回率和 *F*1-score 都有了比较明显的提高.这说明对寄存器采用较精细的标准化规则能够更有效地反映出家族特征,会获得更好的分类结果.

3.2 主题数目的确定

在 LDA 建模过程中,需要预先设定主题的数目,但是如何准确设置主题的数目,这是一个很困难的问题.本文采用了困惑度来评价主题数目对模型的影响.困惑度(perplexity)是一种信息理论的测量方法.若求 A 的困惑度值,则定义为基于 A 的熵的能量(A 可以是一个概率分布或者概率模型):

$$Perplexity(A)=\exp\left\{\frac{-\sum_{d=1}^{M}\lg P(\boldsymbol{W}_d)}{\sum_{d=1}^{M}N_d}\right\}, \quad (12)$$

显而易见随着主题数的增加,困惑度会减小——更多的主题可以把单词更轻松和置信到不同的主题上.但是在主题数极少的时候,困惑度不会随着主题数上升而减小,反而会增加,这是因为在没有到达当前语料库合适的主题数时,大量困惑的样本一直难以被分配到合适的主题上.因此,本文选择拐点作为 LDA 模型的主题数,如图 5 所示:

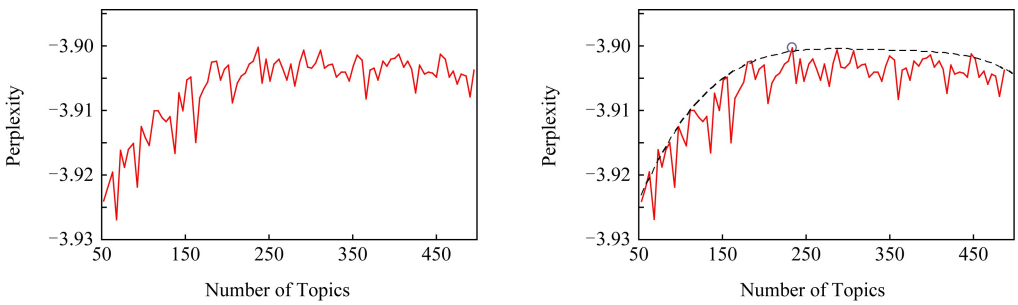


Fig. 5 Perplexity curve of LDA model using different topic numbers

图 5 LDA 模型在不同主题数下困惑度曲线

在确定最优主题数目的过程中,为了加快最优主题选择的效率,本文采取的策略为:主题数目低于 200 时,按照固定步长的递增方法设置主题数目;当超过 200 时采用变化的步长,由大到小地确定主题数目的方法,加快了主题数目确定的速度,减少了一半以上的时间消耗.

3.3 LDA 主题模型的特征描述能力

在 LDA 模型中,为了更明确地表述主题模型的特征描述能力,本文从实验中提取 2 个分布明显的家族——Allaple 与 Adload 进行分析.首先提取 2 个家族中占比最多的几个主题,列出每个主题前十的汇编指令.根据“主题-词”分布,尝试分析其行为.图 6 为 Allaple 家族主题分布图,这里主题数目为 240 个.主题概率分布居前 2 位“主题”中的前十个“词”如表 5 所示.图 7 为 Adload 家族主题分布图,同样地选择主题概率分布居前 2 位的“主题”中前十个“词”如表 6 所示.

在这 2 个家族中可以看到的是,每个具有高拟合度的主题具有各自不同的特点.反映出来的或是在底层上的操作、或是在寄存器上的操作、或是在数据类型上的变化、或是中断的处理、或是函数的调用和返回、或是循环使用、或是权限申请,或是硬件

端口使用等等信息.这些信息具有可读性以及可解释性.

虽然在构造 LDA 模型过程中词包是离散的、无序的形式,可从表面上看损失了汇编指令的上下文信息,但是从表 5、表 6 的分析中也可以看到,本文方法聚集的特征具有隐含的语义信息,这并没有受到"词"序的影响,这就是 LDA 模型的魅力所在.

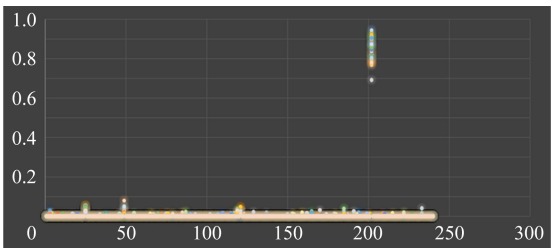


Fig. 6 Probability distribution diagram of Allape family

图 6 Allaple 家族主题分布图

Table 5 Probability Distribution of Top 10 Words in Allaple Family

表 5 Allaple 家族中主题中的词分布

Topic	Top Ten Words	Behavior Description
The 200th Topic	"ire oth", "pus oth", "stc oth", "pop oth", "imu ecx mem oth", "arp mem oth", ".by oth", "dec oth", "inc oth", "xch eax oth"	interrupt return; application for privilege level; exchange the value of register "eax";
The 23th Topic	"pus oth", ".by oth", "inc oth", "pop oth", "dec mem", "dec oth", "jmp oth", "int oth", "cld oth", "std oth"	plus 1; minus 1; loop operartion; string operation;

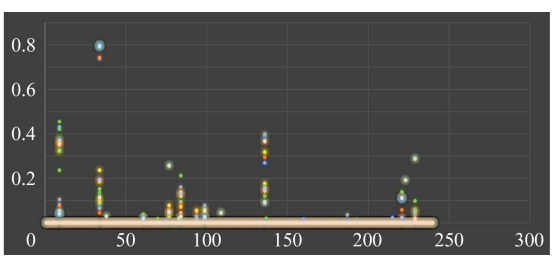


Fig. 7 Probability distribution diagram of Adload family

图 7 Adload 家族主题分布图

Table 6 Probability Distribution of Top 10 Words in Adload Family

表 6 Adload 家族主题中的词分布

Topic	Top Ten Words	Behavior Description
The 7th Topic	"cdq oth", "pus oth", "pop oth", "and ecx mem", ".by oth", "xch eax oth", "mov val oth", "dec oth", "inc oth", "int oth"	extended register higher digits; register; replication and computing operations; plus 1; minus 1;
The 32th Topic	"pop oth", "pus oth", "dec oth", ".by oth", "xch eax oth", "mov val oth", "inc oth", "ret oth", "mov rg8 oth", "int oth"	register assignment operation; function return;

3.4 与其他方法的比较

在 Kaggle 和 CNCERT 数据集上的实验结果验证了本文方法的有效性.下面给出与他人方法的实验对比.

本文在 CNCERT 数据集上完成了其他 3 篇文献方法的实验:

1) 提取恶意代码反汇编文件的操作码序列并转化为点阵图的恶意代码分类实验[26];

2) 将恶意代码二进制可执行文件转化为灰度图像的分类实验[10];

3) 基于恶意代码反汇编文件 Opcode 频度的恶意代码分类实验[4].

对比 3 个方法与本文方法的分类结果,可以看到本文方法与其他分类方法相比具有一致的或者更好的分类精确率,结果如表 7 所示:

Table 7 Classification Results of Many Methods

表 7 多种方法的分类结果

Method	Accuracy
Ref [4]	0.938
Ref [10]	0.902
Ref [26]	0.935
Our Method	0.940

此外,由于 LDA 的无监督学习特性,使得基于主题概率模型的特征提取方法可以为新样本赋予与训练集中的文档相关联的一组不同的主题概率[23].因此,本文方法完全有能力检测新样本,这是文献[4,10,26]不具有的能力.

4 总 结

本文将 LDA 主题模型用于 Windows 平台下的恶意代码分析中,采用 LDA 模型聚类恶意样本的特征,设计恶意代码检测的工作框架,实现恶意样本的分类问题.针对 LDA 模型主题数目不易确定的问题,提出了困惑度的评价方法,并采取了加速策略,大大提高运行效率.同时本文方法还能够检测未知样本.

但是受限于反汇编技术,目前在汇编命令的层级能获得的信息描述更加偏向于底层,如果想获得更抽象的信息可以采用 2 种方法:在静态分析的情况下,通过 IDA 获得 WinAPI;在动态的情况下,可以通过沙箱获得 API 调用的词包信息,预测这些信息应用在本文的方法上依然是可行的,这将是本文接下来的研究内容.

参 考 文 献

[1] Symantec. Internet security threat report [OL]. [2019-03-07]. http://www.symantec.com

[2] Venkataraman S, Blum A, Song D. Limits of learning-based signature generation with adversaries [C/OL] //Proc of the 15th Annual Network and Distributed System Security Symp. San Diego, CA: The Internet Society, 2008 [2019-03-12]. http://intelli-sec.cs.berkeley.edu/papers/18_limits__learning-based.pdf

[3] Moskovitch R, Feher C, Tzachar N, et al. Unknow malcode detection using opcode representation [C] //Proc of the 1st Intelligence and Security Informatics. Berlin: Springer, 2008: 204-215

[4] 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

[5] Kapoor A, Dhavale S. Control flow graph based multiclass malware detection using bi-normal separation [J]. Defence Science Journal, 2016 (66): 138-145

[6] Ashkan S, Babak Y, Hossein R, et al. Malware detection based on mining API calls [C] //Proc of the 2010 ACM Symp on Applied Computing. New York: ACM, 2010: 22-26

[7] Schultz M G, Eskin E, Zadok E, et al. Data mining methods for detection of new malicious executables [C] //Proc of 2011 IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2001: 38-49

[8] Santos I, Devesa J, Brezo F, et al. OPEM: A static-dynamic approach for machine learning based malware detection [C] //Proc of Int Joint Conf CISIS'12-ICEUTE'12-SOCO'12 Special Sessions. Berlin: Springer, 2012: 271-280

[9] Saxe J, Berlin K. Deep neural network based malware detection using two dimensional binary program features [C] //Proc of the 10th Int Conf on Malicious and Unwanted Software. Piscataway, NJ: IEEE, 2015: 11-20

[10] Nataraj L, Karthikeyan S, Jacob G, et al. Malware images: Visualization and automatic classification [C] //Proc of the 8th Int Symp on Visualization for Cyber Security. New York: ACM, 2011. Doi:10.1145/2016904.2016908

[11] Conti G, Bratus S, Shubina A, et al. A visual study of binary fragment types [R]. Las Vegas, Nevada: Black Hat, 2010

[12] Han K S, Lim J H, Kang B J, et al. Malware analysis using visualized images and entropy graphs [J]. International Journal of Information Security, 2015, 14(6): 1-14

[13] Han K S, Kang B J, Im E G. Malware analysis using visualized image matrices [J]. The Scientific World Journal, 2014, 2014: 1-15

[14] Tamersoy A, Roundy K, Horng D C. Guilt by association: Large scale malware detection by mining fille-relation graphs [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 1524-1533

[15] Ye Yanfang, Li Tao, Zhu Shenghuo, et al. Combining file content and file relations for cloud based malware detection [C] //Proc of the 17th ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 222-230

[16] Fan Yujie, Hou Shifu, Zhang Yiming, et al. Gotcha-sky malware! Scorpion: A metagraph2vec based malware detection system [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 253-262

[17] Christodorescu M, Jha S, Kruege C. Mining specifications of malicious behavior [C] //Proc of the 6th Joint Meeting of the European Software Engineering Conf and the ACM SIGSOFT Symp on the Foundations of Software Engineering. New York: ACM, 2008: 5-14

[18] Firdausi I, Lim C, Erwin A, et al. Analysis of machine learning techniques used in behavior-based malware detection [C] //Proc of the 2nd Int Conf on Advances in Computing, Control, and Telecommunication Technologies. Piscataway, NJ: IEEE, 2009: 10-12

[19] Ding Yuxin, Xia Xiaoling, Chen Sheng, et al. A malware detection method based on family behavior graph [J]. Computers & Security, 2018, 73: 73-86

[20] Lee Y S, Lee J U, Soh W Y. Trend of malware detection using deep learning [C] //Proc of the 2nd Int Conf on Education and Multimedia Technology. New York: ACM, 2018: 102-106

[21] Christos H P, Prabhakar R, Hisao T, et al. Latent semantic indexing: A probabilistic analysis [C] //Proc of the 7th ACM SIGACT-SIGMOD-SIGART Symp on Principles of Database Systems. New York: ACM, 1998: 159-168

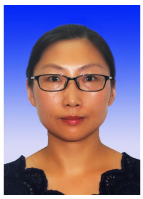
[22] Hofmann T. Probabilistic latent semantic analysis [J]. Uncertainty in Artificial Intelligence, 1999, 41(6): 289-296

[23] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993-1022

[24] Qiao Yanchen, Yun Xiaochun, Tuo Yupeng, et al. Fast reused code tracing method based on simhash and inverted index [J]. Journal on Communications, 2016, 37(11): 104-113 (in Chinese)
(乔延臣, 云晓春, 庹宇鹏, 等. 基于 simhash 与倒排表索引的付用代码块快速溯源方法[J]. 通信学报, 2016, 37(11): 104-113)

[25] Microsoft. Kaggle dataset [OL]. [2018-03-06]. https://www.kaggle.com/c/malware-classification

[26] Liu Yashu, Wang Zhihai, Hou Yueran, et al. Malware visualization and automatic classification with enhanced information density [J]. Journal of Tsinghua University: Philosophy and Social Sciences, 2019, 59(1): 9-14 (in Chinese)
(刘亚姝, 王志海, 侯跃然, 等. 信息密度增强的恶意代码可视化与自动分类方法[J]. 清华大学学报: 自然科学版, 2019, 59(1): 9-14)



Liu Yashu, born in 1977. PhD candidate. Her main research interests include information security, malware detection, machine learning, data ming.

Wang Zhihai, born in 1963. PhD, professor, PhD supervisor. Professor of Beijing Jiaotong University. His main research interests include data mining and business intelligence, machine learning and computation intelligence.

Hou Yueran, born in 1994. Master. His main research interests include information security and data ming.

Yan Hanbing, born in 1975. PhD, professor, PhD supervisor. Professor of the National Computer Network Emergency Response Technical Team/Coordination Center of China. His main research interests include cyber security, image-spam analysis, computer graphics.