一种基于词加权 LDA 模型的恶意文件检测方法

徐建国　王旭阳
（山东科技大学计算机科学与工程学院　山东 青岛 266590）

摘　要　恶意文件中往往含有出现频率较低、但表征能力更好的特征码，传统的方法未能将这一类特征提取出来。针对该问题，提出一种基于词加权 LDA 模型的恶意文件检测方法，该方法通过反汇编对样本进行预处理，采用改进的 KeyGraph 算法（IKG）提取“重点词”，这类词具有更好的特征表征能力，再利用优化的点互信息（OPMI），算出各“重点词”权重，构建词字典，然后将该词加权方法扩展到 LDA 模型，建立 IKG-OPMI-LDA（IOL）模型完成分类，并采用 Gibbs Sampling 进行参数估计。实验结果表明，相较于其他方法，该方法的分类准确率有明显提高，分类效率更好，并且提取的特征具有更高的区分度，与主题相关度更高。

关键词　恶意文件　LDA　IKG　加权模型　文档分类

中图分类号　TP39　　**文献标志码**　A　　DOI：10.3969/j.issn.1000-386x.2024.03.048

A MALICIOUS FILE DETECTION METHOD BASED ON "KEY WORDS" WEIGHTED LDA MODEL

Xu Jianguo　Wang Xuyang
（*School of Computer Science and Engineering*, *Shandong University of Science and Technology*, *Qingdao 266590*, *Shandong*, *China*）

Abstract　Malicious files often contain feature codes that appear less frequently but have better characterization capabilities. Traditional methods have failed to extract this type of feature. In response to this problem, a malicious file detection method based on word weighted LDA model is proposed. The method preprocessed the samples through disassembly, and extracted "key words" by improved KeyGraph algorithm (IKG). This kind of words had better characteristic representation abilities. The optimized point mutual information (OPMI) was used to calculate the weight of each "key word", established a word dictionary. This word weighting method was extended to the LDA model, and the IKG-OPMI-LDA (IOL) model was built to complete the classification. Gibbs Sampling was adopted for parameter estimation. The experimental results show that, compared with other methods, the classification accuracy of this method is significantly improved, the classification efficiency is better, and the extracted features have a higher degree of discrimination and a higher degree of correlation with the topic.

Keywords　Malicious files　LDA　IKG　Weighted model　Document classification

0　引　言

恶意程序指的是一段带有攻击意图的程序，而恶意文件是一种包含恶意程序的可移植可执行代码文件，其格式为 PE 文件，是恶意软件的基本组成部分。近年来，信息安全问题愈发突出，因恶意文件的下载而引起的案件数越来越多，每年增幅可达 50%。据 2020 年国家互联网应急中心发布的《1 月份—6 月份全球互联网安全威胁报告》称，全球日均恶意文件下载次数达到了 156 亿次，几乎每一种物联网终端都会有安全隐患，这对恶意文件的检测方法有了更高的要求。目前的恶意文件检测方法主要是检测文件中的特征码，构建代码特征库，通过比较相同位置的字节[1]，判断样

收稿日期：2021－01－16。2016 年青岛市哲学社会科学规划项目（QDSKL1601121）；2017 年山东省高校人文社会科学研究计划（思想政治教育专题研究）资助经费项目（J17ZZ27）；2018 年山东科技大学研究生科技创新项目（SDKDYC180339）。**徐建国**，副教授，主研领域：网络舆情，信息安全。**王旭阳**，硕士生。

本是否为恶意文件,该方法也被各类反病毒软件采用。但这只能提取出现频率较高的特征,一些出现次数低却能更好体现恶意文件特征的特征码却未能被发现,这会给计算机造成较大的安全隐患。基于此情况,众多学者也发表了许多很有价值的成果。

吴丽娟等[2]提出了一种 Minkowski 距离的加壳 PE 文件识别方法,检测恶意软件是否加壳;杨燕等[3]依照特征频率的相关性,引入变长 N-Gram 特征筛选特征码;Lee 等[4]利用数据挖掘的方法,将代码分成小规模再进行检测;Naeem 等[5]将恶意软件二进制文件转换为灰度图像,并设立本地和全局恶意模式来分类各项特征;这些方法虽然有一定的精确度,但需消耗大量的资源,使得程序在执行时往往耗时较多,降低了效率。

随着机器学习等新技术的出现,恶意文件检测技术也有了更快的发展。Webster 等[6]使用深度学习算法划分了样本的系统调用序列,判定了样本是否存在恶意性;Fan 等[7]以样本、API、机器之间的关系搭建了异构信息网络来检测恶意样本;Ntantogian 等[8]采用回溯的方法,将样本特征进行模糊处理后再分类,也达到了很好的检测效果;李翼宏等[9]基于最大距离和最小估计风险的样本特征选择策略,对主动学习的方法进行了改进,提高了分类准确率;刘亚姝等[10]则结合"困惑度"和变化的步长,利用 LDA 模型获得恶意文件中的"主题分布"以构造样本特征,进而鉴别出恶意样本。这些方法虽然提高了分析速度与精确程度,但未能解决表征力更好的低频(出现频率较低)特征码被高频特征码淹没的问题。

本文在现有的特征提取方法的基础上,提出了一种基于词加权 LDA 模型的恶意文件检测方法。先对恶意文件样本进行预处理,通过反汇编 PE 文件得到汇编文件;再采用 IKG 算法获取表征力更好的"重点词",利用 OPMI 计算各"重点词"权重,构建词字典;最后将词加权扩展至主题模型来进行分类,使用 Gibbs Sampling 估计相关参数,以此搭建一种新的恶意文件检测框架,有效地提高对恶意文件的分类准确率和效率。具体工作流程如图 1 所示。

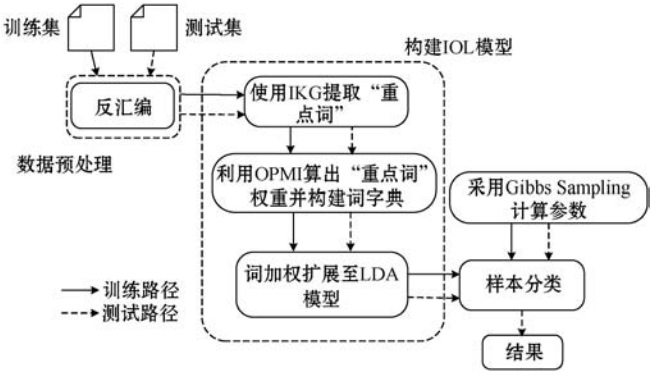

图 1 应用 IOL 模型的恶意文件检测工作流程

1 主题模型

1.1 LDA 模型介绍

概率主题模型是一种统计方法,拥有良好的文档分类能力,它以一种非监督学习的方法对样本中的"词"进行分析,进而挖掘出其中的"文档-主题-词"之间的结构。传统的主题模型有 LSI、PLSA 和 LDA 等。LDA 主题模型(Latent Dirichlet Allocation, LDA)[11]是一种生成模型,可表示为 3 层生成型贝叶斯网,包括词、主题和文档,能够生成"文档-主题"模型。"主题"是样本中"词"的条件概率分布,与"主题"关联度更高的"词",往往出现频率更高。

1.2 LDA 模型定义

在主题模型中,"文档"W 由 N 个词 w_i 组成,即 $W=\{w_1,w_2,\cdots,w_N\}$,w_i 表示第 i 个词;"样本库"D 为 M 篇文档的集合,即 $D=\{d_1,d_2,\cdots,d_M\}$,样本库含有的全部"词"可记为 $W=\{w_1,w_2,\cdots,w_V\}$;"主题"包含于文档中,将与其有关的"词"汇集起来,可记作 $Z=\{z_1,z_2,\cdots,z_Q\}$;该模型中的"主题"、"词"分布都服从狄利克雷(Dirichlet)先验分布。

本文采用的符号与说明如表 1 所示。

表 1 本文使用符号说明表

符号	说明
Q	主题数量
M	文档数量
V	词数
N_d	第 d 篇文档中的词数
θ_d	第 d 篇文档中"文档-主题"的多项式分布,$\boldsymbol{\Theta}=(\theta)_{d=1}^{M}$是 $M\times K$ 矩阵
γ_q	"主题-词"的多项式分布,$\boldsymbol{\phi}=(\gamma_k)_{k=1}^{K}$是 $K\times V$ 矩阵
$Z_{d,n}$	第 d 篇文档中的第 n 个主题
$W_{d,n}$	第 d 篇文档中的第 n 个词
$\boldsymbol{\lambda}$	Dirichlet 分布的超参数,K 维向量
$\boldsymbol{\eta}$	Dirichlet 分布的超参数,V 维向量

LDA 模型概率图如图 2 所示。

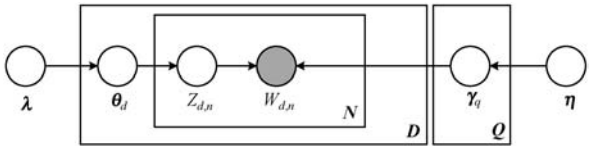

图 2 LDA 模型概率图

该模型是为了找出所有文档的"主题分布"和所有主题中"词的分布"。因为文档中的"词"往往是无

序的,它们彼此互相独立,并且这些词在计算完词频后组成词向量,即“词袋”。“主题”生成“词”的流程不依赖于任何一个文档,故“文档-主题”分布和“主题-词”分布是相互独立的。

2　恶意文件检测方法

2.1　恶意文件样本预处理

恶意样本可以经由逆向工具把 PE 文件(常见的有 DLL、EXE、SYS、OCX 文件等)转换成汇编语言代码,本文使用 Python 包中的 pefile 解析 PE 文件,取得汇编文件,如图 3 所示。

```
.text:01036552    push    ebx
.text:01036553    move    esi, esi
.text:01036556    push    edx
.text:01036553    cmp     edi
.text:01036552    jz      loc_1036A62
.text:0103655D    cmp     short ptr [edp+List], edx
.text:01036553    move    ebx,edi
.text:01036556    push    2                    : IKey
.text:01036552    xor     ecx
.text:0103655E    call    sub_201C31D
.text:01036562    lea     ebx, [esi+IKey]
.text:01036556    call    sub_113C01D
```

图 3　PE 文件反汇编示例

2.2　词权重计算方法

传统的 LDA 模型在分类时,高频词会有更大概率出现,但主题中也会有能更好表现主题的低频词,如 zeus 家族第 7 号主题中的特征词“grap REG,SRC al”(抓取表单、窃取信息的操作),Necurs 家族第 92 号主题中的特征词“movs bml DST”(描述传播恶意信息的行为)、第 57 号主题中的“HLTD FT st. cer”(偷取会话凭证和个人数据的命令),Rammit 家族第 36 号主题中的“JXCZ ser REC opr”(强行控制服务器通信以接受指令)等,这些低频词在传统 LDA 模型下分配到每个主题里的几率较低,而部分高频词表现主题的能力较弱,却有很高的几率被分配到每个主题中,使得模型的主题表现能力变弱。本文先将表达主题能力更好的“重点词”获取出来,再计算“重点词”与样本文档间相关性的值作为权重值。

2.2.1　基于 IKG 的“重点词”提取

KeyGraph(KG)算法[12]源自于建筑框架中将图分割化为群的一种索引方法,框架中的图展示的是文档中各个词之间共同出现的关系,各群集相对应的是思想基础的概念。基于此,本文以 KG 算法为基础,提出了一种 IKG(Improved KeyGraph)算法,依据词与其相应群集间的关系计算出“重点词”,即排序最高的词,这类词当中也包括了和高频词共现的低频词,这些低频词也具备恶意文件的特征。推导过程如下:

(1) 高频词的获取。文档中词频大于指定阈值(TH)的词便是高频词,采用统计方法将其抽取出来,并把选中的高频词在图中以顶点表示,高频词的集合用 C 表示。

$$C = \{ w \in W \cap c\text{-}c_{\text{ounts}} > T_{\text{H}} \} \tag{1}$$

(2) 高频词关联度计算。通过式(2)可计算出两个高频词在同一个句子中具体的共现度,将共现度高出指定阈值的两个高频词连接起来,产生边的集合,此时图中形成一个或多个岛屿。(因实验数据集中包含 10 个家族,为增加计算精确度,本文中所有对数函数底均取为 5。)

$$c_{\text{orrd}} = \sum_{q \in D} \log_5 \frac{2 |w_i \cdot w_j|_q}{|w_i|_q |w_j|_q} \quad w_i, w_j \in C \tag{2}$$

式中:$|w_i \cdot w_j|_q$为词 w_i与词 w_j在句子 q 中共同出现的频数;$|w_i|_q$为词 w_i在句子 q 中的出现频数;$|w_j|_q$为词 w_j在句子 q 中的出现频数;D 为文档库。

(3) 取得“重点词”。IKG 算法定义了一个函数 $crux(w)$,值域为 0 到 1,以注明每个词对其主题所产生的作用大小。此外,本文也设定了两个辅助函数:

$$c\text{-}exist(w,h) = \frac{\sum_{q \in D} \log_5 |w|_q |h - w|_q}{Q} \tag{3}$$

$$embody(h) = \sum_{q \in D} \sum_{w \in Q} \frac{\log_5 |w|_q |h - w|_q}{Q} \tag{4}$$

$$|h - w|_q = \begin{cases} |h|_q - |w|_q & w \in h \\ |h|_q & w \notin h \end{cases} \tag{5}$$

式(3)表示词 w 与群 h 两者共同出现的关系,即 w 作为 h 中的词出现于同句当中的次数($|h - w|_q$)与 w 的词数之间所呈现的关系,值越大,共现关系越强,$|w_q|$为句 q 中 w 的词数,Q 为主题数目;式(4)表示句子里所含有 h 中的词与样本文档中所有词的比例关系,值越小,表示词的出现频率越低;式(5)表示 $|h - w|_q$的取值方式。

由此得出函数 $crux(w)$的公式为:

$$crux(w) = 1 - \prod_{h \subset H} \left(1 - \frac{c\text{-}exist(w,h)}{embody(h)}\right) \tag{6}$$

根据该函数构建高 *crux* 值词集合 KS,将 W 中最大的 r 个 *crux* 值最高的词加入到中,r 取经验值。把在此高 *crux* 值词集合中没有出现于图中的词设为新节点导入到图中。

(4) 高 *crux* 词与高频词的关联度计算如式(7)所示。

$$relev(w_i, w_j) = \sum_{q \in D} \log_5 \frac{2 |w_i \cdot w_j|_q}{|w_i|_q |w_j|_q} \quad w_i \in KS, w_j \in C \tag{7}$$

若图中节点 w_i 与 w_j 无边，且 $relev(w_i,w_j)!=0$，则使用虚线连接起来，得出样本文档的 IKG 图，便于统计高 *crux* 词的相关信息。之后计算图中和群有关的每个节点的 *relev* 值，若该节点 *relev* 值大于指定阈值，则该节点对应的词即为“重点词”。

2.2.2 “重点词”权重值的计算

本文采用优化的点互信息（Optimized Pointwise Mutual Information，OPMI[13]）为评价函数来处理“重点词”的权重值，利用 OPMI 值来表示“重点词” t 与样本文档 d 的相关度。如果“重点词” t 的出现频率在某些文档里较高且在其他文档中较低，它便获得较高的 OPMI 值，公式如下：

$$OPMI(t,d)=\log_5\frac{f_d^{(t)}}{f_t\times f_{d,m}} \tag{8}$$

式中：f_t 为在所有文档中“重点词” t 的出现频数；$f_{d,m}$ 为第 d 篇文档和所有文档的词频比值；$f_d^{(t)}$ 为“重点词” t 在第 d 篇文档中的出现频数。

依据式(8)，“重点词”权重值的计算公式如下：

$$weight(w)=\frac{\sum_{i=1}^{M}OPMI(w,d)}{M}=\frac{\sum_{i=1}^{M}\log_5\dfrac{f_d^{(w)}}{f_w\times f_{d,m}}}{M} \tag{9}$$

在利用 OPMI 计算出权重值后，将“重点词”对应的样本文档编号与权重值储存到词字典中。

2.3 基于词加权的 LDA 模型

2.3.1 词加权 LDA 模型

本文把上节方法与 LDA 模型进行结合，得出一种新的词加权 LDA 模型，即 IKG-OPMI-LDA 模型。此模型由 IKG 算法提取样本文档中的“重点词”，并将这类词记为特征词，构建词字典来存储“重点词”在所属文档中的权重值，从而用此字典将特征词进行加权与归一化处理，以完成对 LDA 模型的改进。

相比于传统的 LDA，新模型中的 word 层采用 IKG 算法对“重点词”和非“重点词”进行标识，这两种词都源于相同的狄利克雷分布，而它们对应生成的概率分布是不同的，使得主题学习的正确性得以提高。IKG-OPMI-LDA（简称为 IOL）模型概率图如图 4 所示。

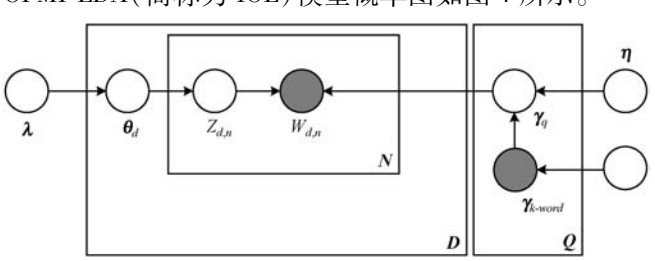

图 4 改进后的 LDA 模型概率图

假定各主题 q 中由一些特征词组成，$w^{(q)}=(w_1,w_2,\cdots,w_n)$，$w_i(i\in[1,n])$，各个词的权重值可从词字典中取得。若该词是“重点词”，其权重值即是字典中“重点词”的权重值；反之，权重值等于 0。得权重向量如下：

$$\boldsymbol{\gamma}_{k\text{-}word}^{(q)}=(s_1,s_2,\cdots,s_n),S_{ij}^{(q)}=\begin{cases}k\text{-}word & w_i\text{为重点词}\\ 0 & \text{其他}\end{cases} \tag{10}$$

每个主题的词概率分布的狄利克雷先验分布为 $\boldsymbol{\eta}^{(q)}=S^{(q)}\times\boldsymbol{\eta}=(\boldsymbol{\eta}_{s_1}(q),\cdots,\boldsymbol{\eta}_{s_n}(q))^{\mathrm{T}}$。

IKG-OPMI-LDA 模型是一种生成式概率图模型，其样本文档库生成步骤如下：

（1）针对整个样本集并依照概率，产生文档主题分布 $\theta_m\sim Dir(\lambda)$。

（2）针对各主题 z，依照概率产生主题特定的词分布 $\phi_{q\mid k\text{-}word}\sim Dir(\eta)$，如果该特定词是“重点词”，则该多项分布向量记为 $\boldsymbol{\phi}_{k\text{-}word}$，反之，多项分布向量为 $\boldsymbol{\phi}_q$。

（3）依据文档中的词对照表，取得文档 d 词数 N_d，N_d 服从泊松分布。

（4）针对文档 d 中各个词 w_n 的生成过程，按照①、②重复执行，迭代 T 次：① 依据词 w_n 和 $Multinomial(\theta_m)$ 判定主题 z，如果该词是“重点词”，便依照词字典修改“重点词”的权重，然后在 θ_m 的 $Multinomial(\theta_m)$ 多项分布里再任意选定一个主题 z；反之，便从 θ_m 的多项分布 $Multinomial(\theta_m)$ 中任意选定一个主题 z。② 在主题 z 的多项条件概率分布中选定一个词 w。

执行完以上流程，就得出了包含 N_d 个词的文档。

2.3.2 模型学习过程

对于 IKG-OPMI-LDA 模型而言，依据样本文档产生的流程逆向推出文档的主题分布，相当于对模型中参数求解的过程。常见的参数估计法有 EM 算法[14]、EP 方法、VEM 方法[15]、VMP 算法、蒙特卡洛采样法[16]（MC）等。本文采用蒙特卡洛采样法中的 Gibbs Sampling 方法[17]对模型中的参数进行推算与改进，此方法基于马尔可夫链，对于统计型的问题可直接进行解决，较其他方法而言有更好的精度与收敛度。由 Gibbs Sampling 得出位于样本文档 d 中第 i 处的词的主题概率分布公式为：

$$p(z_{d,n}\mid z_{\neg d,n},w)\propto\hat{\theta}_{mq}\cdot\hat{\varphi}_{qx}=\frac{n_{q,\neg i}^{(x)}+\eta_x}{\sum_{t=1}^{V}(n_{q,\neg i}^{(x)}+\eta_x)}\times\frac{n_{m,\neg i}^{(q)}+\lambda_q}{\sum_{q=1}^{Q}(n_{m,\neg i}^{(q)}+\lambda_q)} \tag{11}$$

式中：q 为主题数，该采样的流程是一种由文档到主题 $z_i(z_i\in[1,q])$，再由主题的 q 条路径到词的流程；$i=$

(m,n)是二维下标,对应第 m 篇文档的第 n 个词;$\neg i$ 表示取出下标为 i 的词;$n_q^{(t)}$ 表示第 q 个主题生成的词中位置 t 的词个数;n_m^q 表示第 m 篇文档中第 q 个主题生成的词个数。

本文将词权重和模型中主题概率 θ_m、词的主题概率 φ_z 进行结合,将不同的权重分别分配给在不同的主题下的不同特征词,然后对模型产生特征词的概率进行变更,得到式(12)－式(13)。

$$\hat{\phi}_{qx} = \frac{weight(x)n_{q,\neg i}^{(x)} + \eta_x}{\sum_{x=1}^{V}(weight(x)n_{q,\neg i}^{(x)} + \eta_x)} \tag{12}$$

$$\hat{\theta}_{mq} = \frac{\sum_{j=1}^{W} weight(j) \cdot n_{mj,\neg i}^{(q)} + \lambda_q}{\sum_{q=1}^{Q}\left(\sum_{j=1}^{W} weight(j) \cdot n_{mj,\neg i}^{(q)} + \lambda_q\right)} \tag{13}$$

结合式(12)－式(13),推导出新 Gibbs Sampling 公式并改进参数,如式(14)所示。

$$p(z_{d,n} \mid \lambda,\eta,z_{\neg d,n},weight(w)) \propto \frac{weight(x)n_{q,\neg i}^{(x)} + \eta_x}{\sum_{x=1}^{V}(weight(x)n_{q,\neg i}^{(x)} + \eta_x)} \times \frac{\sum_{j=1}^{W} weight(j) \cdot n_{mj,\neg i}^{(q)} + \lambda_q}{\sum_{q=1}^{Q}\left(\sum_{j=1}^{W} weight(j) \cdot n_{mj,\neg i}^{(q)} + \lambda_q\right)} \tag{14}$$

由式(12)－式(14),该模型学习过程的步骤如下:

导入部分:样本文档库、主题数目 Q、超参数。

导出部分:每个词对应的主题号 z。

(1) 先对文档库中的各个词任意匹配一个主题号 z。

(2) 将文档库中的各个词 w 重新取样:如果 w 是“重点词”,便依照式(14)对该词的主题再次取样,并在文档库中完成更新;反之,便按照式(11)对该词的主题再次取样,并于文档库中完成更新。

(3) 反复进行文档库的取样操作,直至 Gibbs Sampling 达到收敛极限。之后计算出文档库“主题-词”的频率矩阵,此矩阵即 IKG-OPMI-LDA 模型。

(4) 依照式(12)－式(13)计算模型中的参数。

3　实验与结果分析

本文选择了网络安全公司 Endgame 于 2018 年发布的 EMBER 数据集进行实验,包含木马、蠕虫、病毒、后门等,一共 10 个家族、40 000 份恶意文件样本,将样本分为训练样本集和测试样本集,比例分别为 75% 和 25%。本文使用该数据集把本文方法和其他学者方法的实验结果作了对比。

实验所用 PC 机为 DELL OptiPlex 7780,主频为 3.1 GHz。采用 Python 3.8,64 位基于提出的 IKG 算法获取“重点词”,利用评估函数计算“重点词”对主题文档的贡献度以建立词字典,通过 Python 实现本文模型,采用 WEKA 工具对分类的效果进行评价。

3.1　特征提取效果对比

本节把基于 IKG 算法的“重点词”提取方法所提取的词,与 LDA、LDA-B[10]、GBDT[18]、DBN[19] 方法选取的特征词进行比较。实验采用的 40 000 份样本中,IKG-OPMI-LDA 模型提取出所有家族的重点词共 342 096 个,在使用 OPMI 算出“重点词”权重之后,有 54 917 个是负值,剔除该部分,得到 287 179 个“重点词”。本节以分类准确率较高的 Nivdort 家族为例,该家族包含 250 个主题且分布明显,选取该家族主题分布居前 2 位的“主题”进行提取,将其中的数据内容完成预处理后,将其他学者的方法和本文方法选取出的词进行比较,结果如表 2 所示。

表 2　各方法提取词对比

Method	Topic	Picked Words(Number of Occurrence)	Behaviour Description
IKG-OPMI-LDA	The 197th Topic	“move ecx, offset_Micr_3”(6), “jnz short loc_20377EF”(5), “arp mem oth”(37),“jmp short opr ip”(7), “and ecx, [esi+6]”(20), “imu ebx me, cl”(28), “apye direction DF_111C02C”(5)	interrupt return; replication and computing operations; extend register higher digits; exchange the value of register “ebx”; application for privilege level
	The 26th Topic	“complement carry CF－0”(7), “set direction flag DF=0”(26), “reload effective address”(6), “call dword ptr sp-12”(32), “and ebx, [esp+6]”(5), “xor al, ofh”(37),“set interrupt IF－0”(5)	register assignment operation; reduce register digits; loop and computing operations; exchange the value of register “esi”; apply of operating level

续表 2

Method	Topic	Picked Words(Number of Occurrence)	Behaviour Description
LDA	The 197th Topic	"push oth"(31), "stc oth (37), "pop dst oth"(28), "doc oth"(39)	plus 1; stos register; loop return; minus 1
	The 26th Topic	"imp oth"(28), "ire oth cl"(27), "pop bst oth"(33), "ret oth"(36)	interrupt return; function return; loop operation; divide 1
LDA-B	The 197th Topic	"cld oth"(29), "xch ebx oth"(35), "jmp mem"(19), ". by oth"(37)	register xlat; computing operations; interrupt return
	The 26th Topic	"imp oth"(28), "xch mem th"(35), "and ecx mem"(33), "ire oth cl"(27)	register assignment; divide 1; interrupt return
GBDT	The 197th Topic	"ret ptr"(26), "stc oth"(36), ". by oth"(39), "imu ebx mem oth"(33)	computing operations; loop return; stos register
	The 26th Topic	"cdq long opr"(31), "ret oth"(33), "jmp oth"(27), "xch eax oth"(35)	higher degits; function return; register assignment
DBN	The 197th Topic	"cld oth"(24), "imu ebx mem oth"(34), "stc oth"(36), "ret ptr"(26), ". by oth"(39)	string operation; exchange the value of register "ebx"; stos register; loop return
	The 26th Topic	"pop oth"(38), "dec oth"(23), "xch ebx mem oth"(36), "jmp oth"(27), "cdq long opr"(31)	loop operation; exchange the value of register "ebx"; higher digits; function return

表 2 中,提取词(Picked Words)后标注的是该词的出现频数(Number of Occurrence)。由表 2 可知,依照 LDA、LDA-B 方法从 Nivdort 家族前 2 位主题中提取的词如"push oth"(31)、"stc oth"(37)、"imp oth"(28)、"ire oth cl"(27),这些词的出现频率均较高,与文件主题的相关度很低;由 GBDT、DBN 方法从 Nivdort 家族前 2 位主题中提取的词也含有"ret ptr"(26)、". by oth"(39)、"cdq long opr"(31)、"jmp oth"(27)等词,这类词对主题的区分效果较差,出现频数也都较大;而 IKG-OPMI-LDA 方法从 Nivdort 家族前 2 位主题里获得的"重点词",包括"jnz short loc_20377EF"(5)、"jmp short opr ip"(7)、"imu ebx me,cl"(28)、"complement carry CF -0"(7)、"call dword ptr sp-12"(32)等词,既含有频数较高的词,也含有频数较低的词,这些词更能代表对应主题,可以作为代表相应文档的"重点词"。

3.2 文档分类效果比较

为验证 IKG-OPMI-LDA 模型在恶意样本文档分类上有更好的效果,本文把 IKG-OPMI-LDA 模型和 LDA 模型、LDA-S-T 模型[20]和 K-LDA 模型[21]的实验结果作比较。分类效果评价标准采用准确率 P(Precision)、召回率 R(Recall)和 F1 值(F1-score)[16,19]。本文模型使用 Gibbs Sampling 方法估算参数,该模型参数设置为:$Q=10$,$\lambda=0.01$,$\eta=0.01$。迭代次数设为 80 至 120 次。IKG-OPMI-LDA 模型、LDA 模型、LDA-S-T 模型和 K-LDA 模型按照相同次数进行迭代,准确率、F1 值和召回率的比较结果如图 5 - 图 7 所示。

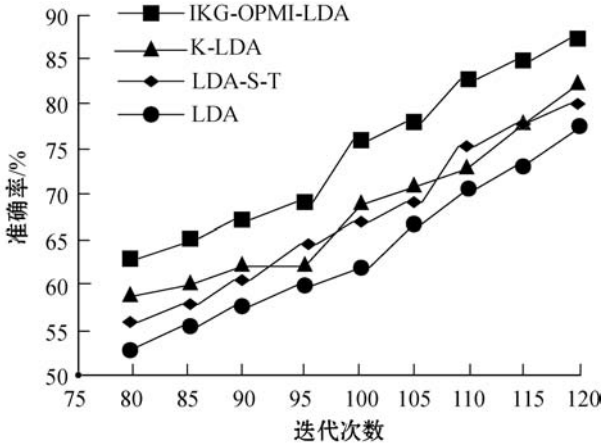

图 5 准确率对比结果

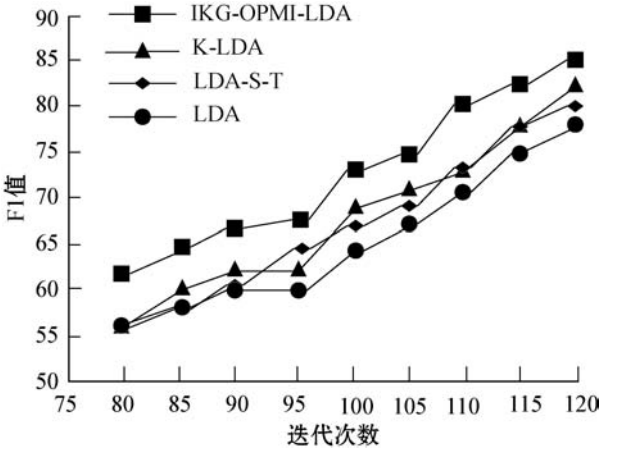

图 6 F1 值对比结果

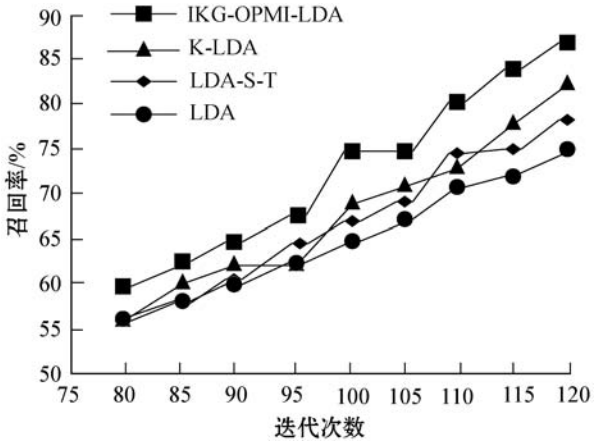

图 7　召回率对比结果

由图 5 – 图 7 可得,IKG-OPMI-LDA 模型在这三个标准上的表现均优于其他三个模型。对于准确率而言,IKG-OPMI-LDA 模型相比于 LDA 模型、LDA-S-T 模型和 K-LDA 模型而言平均提升了 11.92 百分点、7.67 百分点、5.13 百分点;在 F1 值上,IKG-OPMI-LDA 模型相较于其他三个模型分别平均提升了 8.45 百分点、6.96 百分点、4.78 百分点;在召回率方面,IKG-OPMI-LDA 模型相比其他三个模型分别平均提升了 12.18 百分点、8.78 百分点、5.32 百分点。这说明 IKG-OPMI-LDA 模型经由 IKG 算法提取出的“重点词”多数并非为高频特征词,但却对样本文档特征有更好的描述效果,其对分类贡献度较高,在 Gibbs Sampling 中将“重点词”进行加权和归一化,使得“重点词”不会被高频词所淹没,故该模型相较于其他三个模型而言有更高的分类准确性。而且由实验对比结果可知,准确率相同的条件下,IKG-OPMI-LDA 模型的迭代次数相对而言更少、效率更高,相比其他模型进度更快,并得出分类结果。如图 5 所示,当准确率为 70% 时,LDA 模型要迭代 110 次,但 IKG-OPMI-LDA 模型只要完成 97 次迭代流程即可,与 LDA 模型相比少了 13 次;如图 6 所示,F1 值为 67.2 时,LDA 模型要迭代 107 次,但 IKG-OPMI-LDA 模型只要迭代 90 次,相比而言少了 17 次。

IKG-OPMI-LDA 模型计算了“重点词”的权重,使文档特征词的提取效果得以提升。当迭代次数为 100 次时,与 LDA 模型的准确率、召回率、F1 值的对比结果如表 3 所示。

表 3　迭代 100 次的准确率、召回率、F1 值对比

序号	所属家族(测试文档数)	IKG-OPMI-LDA 模型			LDA 模型		
		准确率/%	召回率/%	F1 值	准确率/%	召回率/%	F1 值
1	Necurs (964)	68.32	83.64	79.31	64.22	80.75	75.21
2	Nemucod (1012)	78.64	79.85	79.71	74.42	72.78	73.59

续表 3

序号	所属家族(测试文档数)	IKG-OPMI-LDA 模型			LDA 模型		
		准确率/%	召回率/%	F1 值	准确率/%	召回率/%	F1 值
3	Kelihos (932)	76.43	79.71	77.62	72.23	71.61	71.86
4	Virut (1040)	91.75	80.63	83.54	82.67	79.14	80.62
5	Zeus (988)	87.51	81.59	85.35	84.56	75.91	78.87
6	Nivdort (1036)	98.24	69.46	83.62	95.91	64.35	80.73
7	Mentiger (1012)	85.83	77.54	79.82	80.85	72.81	77.48
8	LMN (960)	79.95	77.23	77.84	75.62	72.45	72.81
9	Rammit (1068)	85.34	87.73	85.97	81.52	82.95	81.82
10	Softpulse (988)	84.52	82.41	82.75	80.37	77.56	79.81
总计	(10 000)	82.62	80.23	81.35	78.55	75.22	77.33

由 LDA 模型从 Nivdort 家族前 2 位主题当中所获取主题词的行为描述来看,包括“plus 1”“loop operation”“minus 1”“function return”等,这些词与主题的关联度较低,证明模型噪声较大。而 IKG-OPMI-LDA 模型得到的主题词行为描述含有“replication and computing operations”“exchange the value of register ‘ebx’”“reduce register digits”“apply of operating level”等,这些描述所对应的词对主题的表现力更强,更适合代表其对应主题。这表明 IKG 算法在选取“重点词”时考虑到了词之间的共现度,进而调整词的权重。由此,可更快地从模型的迭代流程里提取这类“重点词”并记为主题词。将词的权重变更后,这些特征词表达主题的能力得以增强,提升了恶意样本文档分类的准确率。

4　结　语

本文先对数据集预处理,把样本转换为汇编文件,再通过 IKG 算法将表现主题能力更好的“重点词”提取出来,采用 OPMI 算出“重点词”的权重值,并将词加权扩展到 LDA 模型以完成分类,定义 IKG-OPMI-LDA 模型的概率图模型,改进 Gibbs Sampling 采样方法进行参数估计。通过与传统 LDA 模型和其他改进 LDA 模型在同一数据集上的实验结果比较可得,该方法对恶意文件分类有更高的准确率和效率。下一步会对汇

编文件的标准化规则进行钻研,将复杂的数据格式精细化,进一步加快模型的执行速度。

参 考 文 献

[1] 范宇杰,陈黎飞,郭躬德. 软件代码的恶意行为学习与分类[J]. 数据采集与处理,2017,32(3):612 - 620.

[2] 吴丽娟,李阳,梁京章. 一种基于明可夫斯基距离的加壳 PE 文件识别方法[J]. 现代电子技术,2016,39(19):80 - 81.

[3] 杨燕,蒋国平. 基于 N-Gram 的计算机病毒特征码自动提取的改进方法[J]. 计算机科学,2017,44(S2):338 - 341,361.

[4] Lee T H, Kwang-Ho K. A study on detection of small size malicious code using data mining method[J]. Convergence Security Journal,2019,19(1):11 - 17.

[5] Naeem H, Guo B, Naeem M, et al. Identification of malicious code variants based on image visualization[J]. Computer & Electrical Engineering,2019,76:225 - 237.

[6] Webster G, Kolosnjali B, Zarras A, et al. Deep learning for classification of malware system call sequences[C]//29th Australasian Joint Conference,2016:137 - 149.

[7] Fan Y J, Hou S F, Zhang Y M, et al. Gotcha-sky malware Scorpion: A metagraph2vec based malware detection system [C]//24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2018:253 - 262.

[8] Ntantogian C, Poulios G, Karopoulos G, et al. Transforming malicious code to ROP gadgets for antivirus evasion[J]. IET Information Security,2019,13(6):570 - 578.

[9] 李翼宏,刘方正,杜震宇. 一种改进主动学习的恶意代码检测算法[J]. 计算机科学,2019,46(5):92 - 99.

[10] 刘亚姝,王志海,侯跃然,等. 一种基于概率主题模型的恶意代码特征提取方法[J]. 计算机研究与发展,2019,56(11):2339 - 2348.

[11] 陈博,马秀峰. 国内 LDA 模型研究现状可视化分析[J]. 情报探索,2020(11):128 - 134.

[12] Benmalek M, Challal Y, Derhab A. An improved key graph based key management scheme for smart grid AMI systems [C]//IEEE Wireless Communications and Networking Conference,2019:15 - 18.

[13] Sun J C, Yao Y, Xia Y, et al. Exploring the characteristics of acupoints in the treatment of stroke with complex network and point mutual information method[J]. TMR Non-Drug Therapy,2019,2(3):95 - 102.

[14] 张梦琇,周菊玲. 几何分布的参数估计及 EM 算法[J]. 数学的实践与认识,2018,48(20):125 - 128.

[15] Mora D, Rivera G, Velasquez I, et al. A virtual element method for the vibration problem of Kirchhoff plates[J]. ESAIM: Mathematical Modelling and Numerical Analysis, 2018,52(4):1437 - 1456.

[16] 李建伏,巴建军. 基于 MCMC 的 DBSCAN 改进算法[J]. 计算机工程与设计,2020,41(1):122 - 126.

[17] He S M, Shin H S, Tsourdos A. Distributed multiple model joint probabilistic data association with Gibbs sampling-aided implementation[J]. Information Fusion,2020,64:20 - 31.

[18] 胥小波,张文博,何超,等. 一种基于行为集成学习的恶意代码检测方法[J]. 北京邮电大学学报,2019,42(4):89 - 95.

[19] 强晗,郭亚兰,田礼明. 基于深度置信网络的恶意代码检测方法研究[J]. 计算机技术与发展,2019,29(7):93 - 97.

[20] 咎家玮,杨勇. 基于 DOM 树的跨站脚本攻击防御技术研究[J]. 通信与信息技术,2018(3):62 - 67.

[21] 李江华,邱晨. 一种基于元信息的 Android 恶意软件检测方法[J]. 计算机应用研究,2019,36(10):3058 - 3062.

(上接第 312 页)

[8] Lin C, Shen Z, Chen Q, et al. A data integrity verification scheme in mobile cloud computing[J]. Journal of Network and Computer Applications,2017,77:146 - 151.

[9] Anderson R. Two remarks on public key cryptography[C]// ACM Conference on Computer and Communications Security, 1997:135 - 147.

[10] 赵雪娇. 一种具有前向安全的无证书数字签名方案[J]. 信息通信,2019(2):109 - 110.

[11] Song D. Practical forward secure group signature schemes [C]//8th ACM Conference on Computer and Communications Security,2001:225 - 234.

[12] 欧海文,张沙蚌. 基于中国剩余定理的前向安全群签名[J]. 计算机应用,2011,31(S1):98 - 100.

[13] 王硕,程相国,陈亚萌,等. 前向安全的群签名方案[J]. 青岛大学学报(自然科学版),2017(3):38 - 42,47.

[14] 洪璇,张绪霞. 基于中国剩余定理的前向安全群签名方案[J]. 计算机应用研究,2020,37(9):2806 - 2810.

[15] 左黎明,夏萍萍,陈祚松. 一种可证安全的短盲签名方案[J]. 计算机工程,2019,45(12):114 - 118.

[16] 欧海文,雷亚超,王湘南. 一种安全高效的群签名方案[J]. 计算机应用与软件,2020,37(7):309 - 312,328.

[17] 岳笑含,惠明亨,王溪波. 基于群签名的前向安全 VANET 匿名认证协议[J]. 计算机科学,2018,45(S2):382 - 388.

[18] 徐潜,谭成翔,冯俊,等. 基于格的前向安全无证书数字签名方案[J]. 计算机研究与发展,2017,54(7):1510 - 1524.

[19] 王硕,程相国,陈亚萌,等. 基于身份的密钥隔离群签名方案[J]. 计算机工程与应用,2018,54(16):76 - 80.

[20] 王勇兵. 基于离散对数的双向安全签名方案[J]. 青海师范大学学报(自然科学版),2016,32(2):6 - 10.