



ISSN 1001-5965
CODEN BHHDE8

北京航空航天大学 学报

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS



2022-02

Vol.48 No.2

目 次

基于 fastText 算法的行业分类技术	吴震, 冉晓燕, 苗权, 刘纯艳, 张栋, 魏娜 (193)
自适应短文本关键词生成模型	王永剑, 孙亚茹, 杨莹 (199)
基于变分推理的网络舆情传播模式分类	唐红梅, 唐文忠, 李瑞晨, 王衍洋, 王丽宏 (209)
基于启发式遗传算法的模糊测试样本集优化方案	王志华, 王浩帆, 程漫漫 (217)
Android 恶意软件检测低冗余特征选择方法	郝靖伟, 潘丽敏, 李蕊, 杨鹏, 罗森林 (225)
基于随机森林的物联网设备流量分类算法	李锐光, 段鹏宇, 沈蒙, 祝烈煌 (233)
大规模物联网恶意样本分析与分类方法	何清林, 王丽宏, 罗冰, 杨黎斌 (240)
面向异构大数据环境的数据脱敏模型	佟玲玲, 李鹏霄, 段东圣, 任博雅, 李扬曦 (249)
基于异质信息网络的恶意代码检测	刘亚姝, 侯跃然, 严寒冰 (258)
加密文档排序中保序加密算法的最优化选取	张久岭, 黄道超, 沈时军 (266)
电信大数据分析下的时空区域经济可视化应用	李娜, 刘文敏, 孟繁瑞, 刘岩 (273)
一种基于深度学习的恶意代码克隆检测技术	沈元, 严寒冰, 夏春和, 韩志辉 (282)
TLS 密码套件的流量数据随机性分析	郭帅, 程光 (291)
基于情感对象识别和情感规则的微博倾向性分析	
.....	王泽辰, 王树鹏, 孙立远, 张磊, 王勇, 郝冰川 (301)
基于能源分解的用户用电行为模式分析	卢瑞瑞, 于海阳, 杨震, 赖英旭, 杨石松, 周明 (311)
面向智能电网嵌入式设备的网络威胁动态评估方法	吕卓, 郭志民, 陈岑, 莫坚松, 常朝稳 (324)
基于 LIME 的恶意代码对抗样本生成技术	黄天波, 李成扬, 刘永志, 李煜辉, 文伟平 (331)
一种基于攻击距离的对抗样本攻击组筛选方法	刘洪毅, 方字形, 文伟平 (339)
基于抽象汇编指令的恶意软件家族分类方法	李玉, 罗森林, 郝靖伟, 潘丽敏 (348)
基于关键字的海报自动合成系统	关帅鹏, 于海阳, 杨震, 周明, 赖英旭 (356)

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS

Vol. 48 No. 2 (Sum 348) February 2022

CONTENTS

Industry classification technology based on fastText algorithm
..... WU Zhen, RAN Xiaoyan, MIAO Quan, LIU Chunyan, ZHANG Dong, WEI Na (193)

Adaptive short text keyword generation model
..... WANG Yongjian, SUN Yaru, YANG Ying (199)

Classification of network public opinion propagation pattern based on variational reasoning
..... TANG Hongmei, TANG Wenzhong, LI Ruichen, WANG Yanyang, WANG Lihong (209)

Fuzzing testing sample set optimization scheme based on heuristic genetic algorithm
..... WANG Zhihua, WANG Haofan, CHENG Manman (217)

Low redundancy feature selection method for Android malware detection
..... HAO Jingwei, PAN Limin, LI Rui, YANG Peng, LUO Senlin (225)

Traffic classification algorithm of Internet of things devices based on random forest
..... LI Ruiguang, DUAN Pengyu, SHEN Meng, ZHU Liehuang (233)

Large-scale IoT malware analysis and classification method
..... HE Qinglin, WANG Lihong, LUO Bing, YANG Libin (240)

Data masking model for heterogeneous big data environment
..... TONG Lingling, LI Pengxiao, DUAN Dongsheng, REN Boya, LI Yangxi (249)

Malicious code detection based on heterogeneous information network
..... LIU Yashu, HOU Yueran, YAN Hanbing (258)

Optimal choice of order preserving encryption scheme in encrypted document ranking
..... ZHANG Jiuling, HUANG Daochao, SHEN Shijun (266)

Application of space time regional economy visualization based on telecom big data analysis
..... LI Na, LIU Wenmin, MENG Fanrui, LIU Yan (273)

Malicious code clone detection technology based on deep learning
..... SHEN Yuan, YAN Hanbing, XIA Chunhe, HAN Zhihui (282)

Randomness of traffic data in TLS cipher suite
..... GUO Shuai, CHENG Guang (291)

Weibo tendency analysis based on sentimental object recognition and sentimental rules
..... WANG Zechen, WANG Shupeng, SUN Liyuan, ZHANG Lei, WANG Yong, HAO Bingchuan (301)

User electricity consumption behavior mode analysis based on energy decomposition
..... LU Ruirui, YU Haiyang, YANG Zhen, LAI Yingxu, YANG Shisong, ZHOU Ming (311)

A dynamic network threat evaluation method for smart grid embedded devices
..... LYU Zhuo, GUO Zhimin, CHEN Cen, MO Jiansong, CHANG Chaowen (324)

Adversarial sample generation technology of malicious code based on LIME
..... HUANG Tianbo, LI Chengyang, LIU Yongzhi, LI Denghui, WEN Weiping (331)

A method for filtering the attack pairs of adversarial examples based on attack distance
..... LIU Hongyi, FANG Yutong, WEN Weiping (339)

Malware family classification method based on abstract assembly instructions
..... LI Yu, LUO Senlin, HAO Jingwei, PAN Limin (348)

Automatic poster synthesis system based on keywords
..... GUAN Shuaipeng, YU Haiyang, YANG Zhen, ZHOU Ming, LAI Yingxu (356)

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0402

基于 fastText 算法的行业分类技术

吴震¹, 冉晓燕¹, 苗权^{1,2,*}, 刘纯艳³, 张栋¹, 魏娜³

(1. 国家计算机网络应急技术处理协调中心, 北京 100029;
2. 国家计算机网络应急技术处理协调中心北京分中心, 北京 100055; 3. 长城计算机软件与系统有限公司, 北京 100190)

摘 要: 随着中国经济的高速发展和技术创新能力的不断提升,高效的组织、分类信息是提供个性化行业管理和跟踪分析的基础。根据行业信息特点和发展规律,提出了一种基于 fastText 算法的行业分类模型。首先,构建行业分类关键词库,通过特征词库进行分词和权重计算。然后,构建分类器模型,实现中文行业的自动分类。最后,实验选取了 80 000 个包含企业经营范围、企业信息、舆论信息的测试文档,结果表明,所提模型结果高于 Bayes、决策树、KNN 等分类算法,取得了较好的应用效果。

关 键 词: 自然语言处理; 行业分类; fastText 算法; 关键词; 语法模型

中图分类号: TP391.1

文献标志码: A **文章编号:** 1001-5965(2022)02-0193-06

中国的企业数量众多,行业类别广泛。对于广大企业进行行业分类是十分必要的,能够进一步解释特定行业所处的发展阶段及其在国民经济中的地位。以《国民经济行业分类》^[1](GB/T 4754—2017)为例,将行业分为了 97 个大类、473 个中类、1 380 个小类。近年来,网信企业规模快速增加,随着基于深度学习的文本分类技术^[2-4]的提升,如何利用企业数据信息进行行业智能判别和分类具有重要意义,同时对于全国企业的智能分类具有借鉴和参考作用。

据不完全统计,中国网信行业企业数量不低于 200 万家,既分布于互联网、信息设备制造、信息传输、信息技术服务等国民经济行业范围,又同时存在于电商、金融、教育、医疗甚至是大数据、人工智能等新兴和前沿领域。如何对网信行业进行行业分类,是做好网信企业统计分析工作的前提和基础,进而整体掌握当前中国网信行业及重点网信企业发展情况,及时发现运行过程中表现出的苗头性、倾向性、潜在性问题,并提出相对应的政策建议。

目前,进行行业分类的过程中,往往存在覆盖面积小、方式单一、时效性不足等问题,仅依靠传统的人工识别方式无法满足需求,亟须一种精确、高效的识别方式。随着人工智能技术的不断发展和进步,针对文本分类问题的研究也出现多元化的发展。文本分类一般包括 3 个步骤:①文本向量化;②根据算法提取文本特征向量;③使用分类器对特征向量进行分类判断。因此,词表示的好坏将直接影响分类器的结果,如何合理而有效地向量化文本通常是文本分类的关键技术。

词向量最初的表示形式是 one-hot,该方法的缺点是:①词与词间的语义视为相互独立,无法表达词之间的语义信息。②每种表达每次只有 1 个位于高位,其余均位于低位,可能导致维度爆炸,这造成了 one-hot 表示具有高稀疏的特性。TF-IDF 算法解决了以上部分问题,其用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。算法主要思想是:字词在文件中出现的次数增多,其重要性也相应增加,但会与语

收稿日期: 2020-08-09; 录用日期: 2020-09-25; 网络出版时间: 2021-03-10 10:14
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210309.1320.007.html

* 通信作者。E-mail: miaoq@cert.org.cn

引用格式: 吴震, 冉晓燕, 苗权, 等. 基于 fastText 算法的行业分类技术[J]. 北京航空航天大学学报, 2022, 48(2): 193-198.
WU Z, RAN X Y, MIAO Q, et al. Industry classification technology based on fastText algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 193-198 (in Chinese).

料库中出现的频率反比例下降。

2013 年,Mikolov 等^[5]提出了 word2vec 模型,相对于以上方法,该模型能够产生稠密的向量,避免维度爆炸,在语法和语义的相似性方面有很大的提高。

传统文本分类算法有 Bayes、支持向量机(SVM)、KNN、Logistic 回归等。徐军等^[6]使用 Naïve Bayes 等机器学习方法进行新闻的情感自动分类,但是 Naïve Bayes 对缺失数据不敏感,需要假设先验模型,从而导致预测效果不佳。2014 年,Kim^[7]提出了 TextCNN 网络,验证了使用预训练的词向量比使用随机初始化的词向量能获得更好的效果,而使用预训练的词向量初始化,并在具体任务中进行词向量的 fine-tuning 可能会得到更好的效果,但是由于卷积神经网络(CNN)更关注局部特征,并且大的卷积核会导致计算量暴增,不利于模型深度的增加,计算性能也会降低。2017 年,Vaswani 等^[8]完全抛弃了循环神经网络(RNN)和 CNN 等网络结构,首次提出将注意力机制应用在 NLP 领域,注意力机制也成为了大家近期的研究热点。2019 年,Devlin 等^[9]提出了一种新的语言表征模型 BERT,这是一种基于 Transformer 的双向编码器。

本文提出了一种融合多元数据的 fastText 算法^[10],用于网信行业文本分类。该方法在 fast-Text 算法基础上,对 n 元语法模型(n -gram)处理后的输入词序列进行权重计算。同时,融合网信企业信息不同维度的短文本数据,并进行加权计算,实现快速进行网信企业分类的目标。在网信企业数据集上进行实验,分类精确率达到 83.1%。

1 相关工作

1.1 数据收集

本文实验通过获取企业经营范围进行企业识别分类。主要获取方式包括:通过政府数据公开平台中提供的公开 API 接口获取(如北京市政务数据资源网、广东省政府数据统一开放平台、上海市政府数据服务网)企业名称、企业法人、经营范围等信息;通过互联网采集的方式获取企业名称及相关信息。

1.2 文本预处理

文本预处理^[11]是对非结构化企业经营范围、舆论等文本信息进行分析,经过过滤、提炼、选定等一系列自然语言处理操作^[12]后,通过深度神经网络方法快速自动识别、划分文本中的经营范围词,并对智能分词结果文本进行展示。在网信企业分类的工程应用中,通过数据洗清、分词、词性标注、去停用词 4 个方面来完成语料的预处理工作。

2 网信企业分类

2.1 分类原则

以《国民经济行业分类》^[1](GB/T 4754—2017)和《电子信息产业行业分类目录》为基础,对国民经济行业分类中符合“网信”范畴的相关活动进行再归类。分类结果考虑中国宏观经济发展、网信事业发展进程的现实情况,以体现阶段性发展特征。

2.2 分类结果

共分为四大类:信息设备制造、信息传输服务、软件和信息技术服务业、互联网和相关服务,一级、二级、三级目录分类如表 1 所示。

表 1 行业分类结果

Table 1 Industry classification results

一级分类	二级分类	三级分类
信息 设备 制造	计算机 制造	计算机整机制造
		计算机零部件制造
		计算机外围设备制造
		工业控制计算机及系统制造
		信息安全设备制造
		其他计算机制造
	通信设备 制造	通信系统设备制造
		通信终端设备制造
		通信配套产品及其他通信设备制造
	广播电视 设备制造	广播电视节目制作及发射设备制造
		广播电视接收设备制造
		广播电视专用配件制造
		专业音响设备制造
	雷达及配套 设备制造	应用电视设备及其他广播电视设备制造
		雷达及配套设备制造
	非专业视听 设备制造	电视机制造
		音响设备制造
		影视录放设备制造
		家用电子电器主要配套件
		可穿戴智能设备制造
	智能消费 设备制造	智能车载设备制造
		智能无人飞行器制造
		服务消费机器人制造
		其他智能消费设备制造
	电子器件 制造	电子真空器件制造
		半导体分立器件制造
		集成电路制造
显示器件制造		
半导体照明器件制造		
光电子器件制造		
电子元件及 电子专用 材料制造	其他电子器件制造	
	电阻电容电感元件制造	
	电子电路制造	
	敏感元件及传感器制造	
	电声器件及零件制造	
	电子专用材料制造	
	其他电子元件制造	

续表		
一级分类	二级分类	三级分类
信息设备制造	电子信息机电产品制造	电子微电机制造
		电子电线电缆制造
		光纤、光缆制造
		电池制造
	专用仪器仪表制造	电子测量仪器制造
	其他电子设备制造	其他电子设备制造
信息传输服务	电信	固定电信服务
		移动电信服务
		其他电信服务
	广播电视传输服务	有线广播电视传输服务
		无线广播电视传输服务
	卫星传输服务	广播电视卫星传输服务
		其他卫星传输服务
软件和信息技术服务业	软件开发	基础软件开发
		支撑软件开发
		应用软件开发
		其他软件开发
		集成电路设计
	信息系统集成和物联网技术服务	信息系统集成服务
		物联网技术服务
	运行维护服务	运行维护服务
	信息处理和存储支持服务	信息处理和存储支持服务
	信息技术咨询服务	信息技术咨询服务
	数字内容服务	地理遥感信息服务
		动漫、游戏数字内容服务
		其他数字内容服务
	其他信息技术服务业	呼叫中心
		其他未列明信息技术服务业
互联网和相关服务	互联网接入及相关服务	互联网接入及相关服务
	互联网信息服务	互联网搜索服务
		互联网游戏服务
		互联网其他信息服务
	互联网平台	互联网生产服务平台
		互联网生活服务平台
		互联网科技创新平台
		互联网公共服务平台
		其他互联网平台
	互联网安全服务	互联网安全服务
	互联网数据服务	互联网数据服务
	其他互联网服务	其他互联网服务

2.3 企业分类技术方法

2.3.1 分类识别过程

本文实验采用自然语言预处理技术,主要针对网信行业企业的经营范围进行分段、分句,并对每一句进行分词,根据分词与关键词比对的结果,

实现对网信行业统计分类和创新性分类的自动识别与发现。流程如图 1 所示。

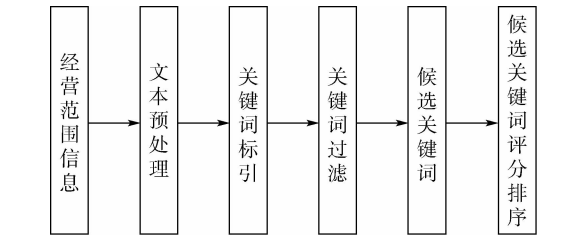


图 1 分类识别流程

Fig. 1 Flowchart of classification and identification

2.3.2 基准分类算法

实验分别采取了 Naïve Bayes、决策树、KNN、TextCNN 等作为基准方法。

2.3.3 fastText 算法

综合考虑经营范围数据的特征,本文基于 fastText 算法进行行业分类识别。fastText 是 facebook 开源的一个词向量与文本分类工具^[13-14],典型应用场景是“带监督的文本分类问题”,结合了自然语言处理和机器学习中最成功的理念。该算法只有隐层和输出层,模型简单,训练速度快,准确率较高。在普通的 CPU 上可以实现分钟级别的训练,比深度模型的训练要快几个数量级,同时,在多个标准的测试数据集上,fastText 在文本分类的准确率上,与现有的一些深度学习的方法效果相当或接近。fastText 使用词袋及 *n*-gram 袋表征语句,使用子词(subword)信息,并通过隐藏表征在类别间共享信息。另外,采用了一个 Softmax 层级(利用了类别不平衡分布的优势)来加速运算。

2.3.4 层次 Softmax

对于有大量类别的数据集,fastText 使用了一个分层分类器^[15](而非扁平式架构)。不同的类别被整合进树形结构中。在某些文本分类任务中类别很多,计算线性分类器的复杂度高。fastText 模型使用层次 Softmax,对标签进行编码,能够极大地缩小模型预测目标的数量。

层次 Softmax 如图 2 所示。

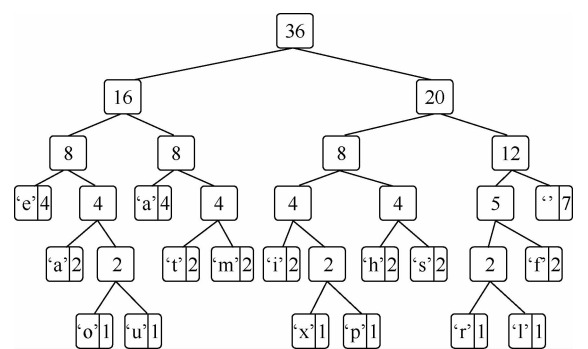


图 2 层次 Softmax

Fig. 2 Hierarchical Softmax

2.3.5 *n*-gram 子词特征

fastText 在用于网信企业名称分类时,使用的特征是词袋模型,同时,加入了 *n*-gram 特征,用于解决词语之间顺序的分类问题。每个词被看做是 *n*-gram 字母串包。为了区分前后缀情况,“<”、“>”符号被加到了词的前后端。除了词的子串外,词本身也被包含进了 *n*-gram 字母串包。以 where 为例,*n* = 3 的情况下,其子串分别为 < wh, whe, her, ere, re >, 以及其本身。

2.3.6 fastText 算法模型

fastText 模型架构^[16]如图 3 所示。

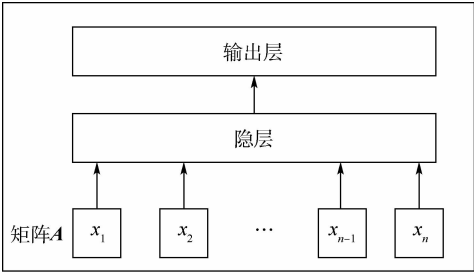


图 3 fastText 模型架构
Fig. 3 Architecture of fastText model

图 3 展示了单个隐层的简单模型^[17]。第一个权重矩阵 *A* 可以被视作某个句子的词查找表。将词表示平均成一个文本表示。文本表示是一个隐藏变量,然后将其送入一个线性分类器。该模型将一系列单词作为输入并产生一个预定义类的概率分布。本文使用 Softmax 函数 *f* 来计算预定义类的概率分布,对于一组包含 *N* 个文档的文档集,fastText 模型目标是使式(1)最小化。

$$\frac{1}{N} \sum_{n=1}^N y_n \ln(f(\mathbf{B} \mathbf{A} x_n)) \tag{1}$$

式中:*n* 为样本个数;*x_n* 为第 *n* 个文档特征的标准包(第 *n* 个样本的归一化特征,每个特征是词向量的平均值);*y_n* 为第 *n* 个样本对应的类别;*f* 为损失函数 Softmax;*A* 为权重矩阵(构建词,embedding);*B* 为权重矩阵(隐层到输出层)。因预测语料更偏向词语的组合,本实验中融入 TF-IDF 的先验特征,经验证显著提升了分类效果。

3 实验与分析

3.1 评价方式

本文采用精确率 *P*、召回率 *R*、*F₁* 值进行评估,具体计算公式如下:

$$P = \frac{A}{C} \tag{2}$$

$$R = \frac{A}{B} \tag{3}$$

$$F_1 = \frac{2PR}{P + R} \tag{4}$$

式中:*A* 为预测正确的正样本数;*B* 为测试数据集中各个类的数目;*C* 为预测结果中各个类的数量;*F₁* 为预测错位的负样本数。

3.2 实验数据

本文实验数据来源于政府信息公开 API 接口及互联网爬取数据。其中,公共 API(北京市政务数据资源网、广东省政府数据统一开放平台、上海市政府数据服务网)和互联网爬取结果获取企业信息(如企业名称、企业法人、企业经营地址等企业信息)。获取到的数据在经过清洗整合后,保存企业名称等结构化数据,通过企业基本信息分析,获取对应企业经营范围。共使用了 10 万条数据,包括计算机制造、通信设备制造、广播电视设备制造、雷达及配套设备制造、非专业视听设备制造、智能消费设备制造、电子器件制造、电子元件及电子专用材料制造、电子信息机电产品制造、专用仪器仪表制造、其他电子设备制造等 31 个企业分类。以上数据根据获取平台划分为 5 个数据集 DS1 ~ DS5,以及 1 个测试集 Test1,如表 2 所示。

表 2 文本数量数据集样本

Table 2 Text quantity dataset sample						
类别	DS1	DS2	DS3	DS4	DS5	Test1
1	3 000	2 000	5 000	2 000	10 000	2 000
2	3 000	2 000	5 000	2 000	10 000	2 000
3	3 000	2 000	5 000	2 000	10 000	2 000
4	3 000	2 000	5 000	2 000	10 000	2 000
5	3 000	2 000	5 000	2 000	10 000	2 000
6	3 000	2 000	5 000	2 000	10 000	2 000
7	3 000	2 000	5 000	2 000	10 000	2 000
8	3 000	2 000	5 000	2 000	10 000	2 000
9	3 000	2 000	5 000	2 000	10 000	2 000
10	3 000	2 000	5 000	2 000	10 000	2 000
11	3 000	2 000	5 000	2 000	10 000	2 000
12	3 000	2 000	5 000	2 000	10 000	2 000
13	3 000	2 000	5 000	2 000	10 000	2 000
14	3 000	2 000	5 000	2 000	10 000	2 000
15	3 000	2 000	5 000	2 000	10 000	2 000
16	3 000	2 000	5 000	2 000	10 000	2 000
17	3 000	2 000	5 000	2 000	10 000	2 000
18	3 000	2 000	5 000	2 000	10 000	2 000
19	3 000	2 000	5 000	2 000	10 000	2 000
20	3 000	2 000	5 000	2 000	10 000	2 000
21	3 000	2 000	5 000	2 000	10 000	2 000
22	3 000	2 000	5 000	2 000	10 000	2 000
23	3 000	2 000	5 000	2 000	10 000	2 000
24	3 000	2 000	5 000	2 000	10 000	2 000
25	3 000	2 000	5 000	2 000	10 000	2 000
26	3 000	2 000	5 000	2 000	10 000	2 000
27	3 000	2 000	5 000	2 000	10 000	2 000
28	3 000	2 000	5 000	2 000	10 000	2 000
29	3 000	2 000	5 000	2 000	10 000	2 000
30	3 000	2 000	5 000	2 000	10 000	2 000
31	3 000	2 000	5 000	2 000	10 000	2 000

3.3 实验环境

实验机器配置:操作系统 Ubuntu 18. 04. 1 LTS,Intel(R)Xeon(R) Gold 6130 CPU@ 2. 10 GHz, fastText 在 gcc 7. 3. 0 上编译。

3.4 实验结果分析

为了验证 fastText 对企业分类的效果,本文使用 DS1 做训练集,使用 Test1 做测试集, n -gram 使用的默认值为 1,维度使用默认值为 200,分类的准确率的宏平均值及所用时间如表 3 所示。

依据表 3,本文实验基于 fastText 算法在 8 s 内完成 80 000 条数据的企业分类,精准率平均值达到 83. 1%,召回率平均值达到了 80. 5%, F_1 值平均值达到了 81. 8%,在网信行业分类方面达到了良好的效果。

表 3 数据集的分类结果				
Table 3 Classification result of dataset				
模型	数据集	精确率	召回率	F_1
Bayes	DS1	73. 6	71. 8	72. 7
	DS2	74. 1	70. 2	72. 1
	DS3	75. 2	73. 8	74. 5
	DS4	69. 9	68. 5	69. 2
	DS5	70. 3	67. 6	68. 9
决策树	DS1	78. 0	72. 4	75. 1
	DS2	76. 8	75. 2	76. 0
	DS3	77. 7	74. 9	76. 3
	DS4	75. 8	72. 1	73. 9
	DS5	72. 1	70. 2	71. 1
KNN	DS1	76. 2	72. 1	74. 1
	DS2	75. 4	70. 3	72. 8
	DS3	70. 1	68. 9	69. 5
	DS4	72. 5	70. 8	71. 6
	DS5	72. 4	69. 8	71. 1
TextCNN	DS1	80. 1	75. 2	77. 6
	DS2	79. 2	75. 0	77. 0
	DS3	81. 1	76. 2	78. 6
	DS4	77. 9	74. 3	76. 1
	DS5	80. 5	78. 2	79. 3
fastText	DS1	82. 2	80. 1	81. 14
	DS2	81. 3	79. 8	80. 5
	DS3	81. 6	79. 9	80. 7
	DS4	82. 5	80. 0	81. 2
	DS5	84. 0	81. 1	82. 5
fastText_TF_IDF	DS1	83. 73	81. 91	82. 81(+ 1. 67)
	DS2	82. 6	79. 1	80. 81(+ 0. 31)
	DS3	83. 1	80. 5	81. 78(+ 1. 08)
	DS4	82. 1	79. 1	80. 57
	DS5	84. 2	82. 0	83. 08(+ 0. 58)
均值		83. 1	80. 5	81. 8

4 结 束 语

针对目前企业分类效率及准确率较低的问题,

本文以网信行业为例,使用 fastText 文本分类算法,探讨其在行业分类的适用性,经实验表明,基于 fastText 的行业分类识别准确率高(可达到 82% 以上),分类速度快,能够满足现有的行业分类需求。

参考文献 (References)

[1] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. 国民经济行业分类:GB/T 4754—2017 [S]. 北京:中国标准出版社,2017.
General Administration of Quality Supervision, Inspection and Quarantine of the People’s Republic of China, Standardization Administration. Industrial classification for national economic activities:GB/T 4754—2017 [S]. Beijing: Standards Press of China,2017(in Chinese).

[2] 石文娟,龙舜,云飞. 基于背景学习的迭代式文本分类框架 [J]. 计算机工程与应用,2015,51(9):991-994.
SHI W J, LONG S, YUN F. Iterative text classification framework based on background learning [J]. Computer Engineering and Applications,2015,51(9):991-994(in Chinese).

[3] HEPBURN J. Universal language model fine-tuning for patent classification [C] // Proceedings of Australasian Language Technology Association Workshop,2018:93-96.

[4] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information [EB/OL]. (2017-06-19) [2020-08-01]. <https://arxiv.org/abs/1607.04606v2>.

[5] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word represent in vector space [EB/OL]. (2013-09-07) [2020-08-01]. <https://arxiv.org/abs/1301.3781>.

[6] 徐军,丁宇新,王晓龙. 使用机器学习方法进行新闻的情感自动分类 [J]. 中文信息学报,2007,21(6):95-100.
XU J, DING Y X, WANG X L. Automatic sentiment classification of news based on machine learning [J]. Journal of Chinese Information Processing,2007,21(6):95-100(in Chinese).

[7] KIM Y. Convolutional neural networks for sentence classification [EB/OL]. (2014-09-03) [2020-08-01]. <https://arxiv.org/abs/1408.5882>.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. (2017-12-06) [2020-08-01]. <https://arxiv.org/abs/1706.03762>.

[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24) [2020-08-01]. <https://arxiv.org/abs/1810.04805>.

[10] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification [EB/OL]. (2016-08-09) [2020-08-01]. <https://arxiv.org/abs/1607.01759v3>.

[11] 闫琰. 基于深度学习的文本表示与分类方法研究 [D]. 北京:北京科技大学,2016.
YAN Y. Research on text representation and classification method based on deep learning [D]. Beijing: University of Science and Technology Beijing,2016(in Chinese).

[12] YESSENALINA A, CARDIE C. Compositional matrix-space

models for sentiment analysis[C] // Conference on Empirical Methods in Natural Language Processing, 2011:172-182.

[13] 代令令,蒋侃. 基于 fastText 的中文文本分类[J]. 计算机与现代化, 2018, 5:37-38.

DAI L L,JIANG K. Chinese text classification based on fastText [J]. Computer and Modernization, 2018, 5: 37-38 (in Chinese).

[14] 孙茂松,李景阳,郭志茆,等. THUCTC:一个高效的中文文本分类工具包[EB/OL]. (2016-01-25) [2020-08-01]. <http://thuctc.thunlp.org>.

SUN M S,LI J Y,GUO Z P,et al. THUCTC:An efficient Chinese text classifier[EB/OL]. (2016-01-25) [2020-08-01]. <http://thuctc.thunlp.org>(in Chinese).

[15] 陆彦婷,陆建峰,杨静宇. 层次分类方法综述[J]. 模式识别与人工智能, 2013, 26(12):1130-1139.

LU Y T,LU J F,YANG J Y. A survey of hierarchical classification methods [J]. Pattern Recognition and Artificial Intelligence, 2013, 26(12):1130-1139(in Chinese).

[16] LE Q,MIKLOV T. Distributed representations of sentences and documents[EB/OL]. (2014-05-22) [2020-08-01]. <https://arxiv.org/abs/1405.4053>.

[17] MANSUR M, PEI W, CHANG B. Feature-based neural language model and Chinese word segmentation[C] // IJCNLP, 2013:1271-1277.

Industry classification technology based on fastText algorithm

WU Zhen¹, RAN Xiaoyan¹, MIAO Quan^{1,2,*}, LIU Chunyan³, ZHANG Dong¹, WEI Na³

- (1. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China;
2. Beijing Branch of National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100055, China;
3. Great Wall Computer Software & System Inc., Beijing 100190, China)

Abstract: With the rapid development of China’s economy and the continuous improvement of technological innovation ability, efficient organization and classification information is the basis of providing personalized industry management and tracking analysis. According to the characteristics of industry information and the law of development, a Chinese industry classification model based on fastText is proposed in this paper. First, the keyword database of industry classification is constructed, then word segmentation and weight calculation are carried out by feature lexicon, and finally the classifier model is constructed to realize the automatic classification of industry. In the experiment, 80 000 test documents including business scope, enterprise information and public opinion information were selected. The results show that the classification accuracy of the proposed model is higher than that of Bayes, decision tree, KNN and other classification algorithms. Thus, the proposed model works well in the application.

Keywords: natural language processing; industry classification; fastText algorithm; keywords; grammar model

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0601

自适应短文本关键词生成模型

王永剑, 孙亚茹, 杨莹*

(公安部第三研究所, 上海 201204)

摘 要: 关键词抽取对文本处理影响较大,其识别的准确度及流畅程度是任务的关键。为有效缓解短文本关键词提取过程中词划分不准确、关键词与文本主题不匹配、多语言混合等难题,提出了一种基于图到序列学习模型的自适应短文本关键词生成模型 ADGCN。模型采用图神经网络与注意力机制相结合的方式作为对文本信息特征提取的编码框架,针对词的位置特征和语境特征编码,解决了短文本结构不规律和词之间存在关联复杂信息的问题。同时采用了一种线性解码方案,生成了可解释的关键词。在解决问题的过程中,从某社交平台收集并公布了一个标签数据集,其包括社交平台发文本和话题标签。实验中,从用户需求角度出发对模型结果的相关性、信息量、连贯性进行评估和分析,所提模型不仅可以生成符合短文本主题的关键词,还可以有效缓解数据扰动对模型的影响。所提模型在公开数据集 KP20k 上仍表现良好,具有较好的可移植性

关 键 词: 关键词提取; 关键词生成; 图神经网络; 注意力机制; 主题模型

中图分类号: TP391

文献标志码: A **文章编号:** 1001-5965(2022)02-0199-10

关键词是文档内容的紧凑表示,由一个或多个字组成^[1-2]。对短文本进行关键词匹配可以有效表达文本主题,关键词识别的准确度对文本分类、文本推荐及文本搜索存在着较大的影响,是文本挖掘和文本分析领域中一项重要的基础工作^[3]。正确提取文本的关键词,有助于对文本的主题进行全面的分析和研究,亦可以对文本做出很好的决策。但因短文本在内容结构上表达灵活和主题风格各异等特点,使得从短文本中提取关键词成为了一项具有挑战性的任务。

分析社交平台发文数据最重要的任务之一是关键字提取。由于社交平台发文本较短,字数限制 140 个字符,内容多样性、非正式性、语法错误、流行语、俚语及实时内容的生成速度等问题,需要一种有效的技术来提取有用的关键词^[4]。传统的关键词提取方法主要是基于机器学习^[5-7]

或图模型^[1,8-9]。机器学习方法通常采用特征值量化的方式从预处理后构建的候选集中得到关键词,如基于频率特征的 TF-IDF 模型、挖掘词句隐藏信息的 LAD 和 HMM 模型。词划分的准确度会对这些模型造成较大的扰动,并且存在语义缺失等问题。图模型对文本构建网络图,词为节点,边为词间的关系。在对图进行分析的过程中,通过评估节点重要性来选出关键词,这增强了节点间的语义关联,如基于拓扑结构的 TextRank 和在 TextRank 的基础上改进的 NE-Rank^[10]。NE-Rank 是一种节点-边的加权方法,通过构造主题关键词特征来有效改进关键词提取。尽管基于机器学习和图模型的方法取得了良好的效果,但是会受特征选择和人工定义规则的影响。近年来,随着深度学习技术的出现,利用深度学习算法自动学习特征可以提高许多任务的性能。采用卷积神经网络

收稿日期: 2020-10-23; 录用日期: 2020-11-06; 网络出版时间: 2021-02-18 08:57
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210215.1310.003.html
* 通信作者。E-mail: yangying@mcst.org.cn

引用格式: 王永剑, 孙亚茹, 杨莹. 自适应短文本关键词生成模型[J]. 北京航空航天大学学报, 2022, 48(2): 199-208.
WANG Y J, SUN Y R, YANG Y. Adaptive short text keyword generation model[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 199-208 (in Chinese).

络在文本字符级上建模,学习文本和单词间的分布矢量表示,以捕获有关单词的形态信息,有效缓解了未知单词的问题^[11]。然而卷积神经网络无法捕获长距离依赖,获取到的是文本局部特征。杨丹浩等^[12]采用序列标注模型从字粒度和词粒度角度结合注意力机制^[13-14]提取文本特征,有效改进中文文本的关键词提取。序列标注模型在文本处理领域取得了较好的进展,基于深度 Bi-LSTM 的 ELMo^[15]可以学习到词在上下文中复杂的用法,但是序列模型当前步的计算依赖前一步的计算,无法考虑另一个方向的数据。图神经网络(graph neural network, GNN)的出现缓解了深度学习模型处理文本数据结构的一系列问题。图神经网络将词、句子或词句搭配表示成图中的节点,节点的表征依赖与其相邻的节点,边表示节点间不同类型的关系,在图结构上应用神经网络可以直接获取节点依赖信息。但是基于有监督学习的模型依赖标签数据,存在模型移植性能差、不能有效发现新词等问题。综上,如何使模型可以正确划分词并能够识别出可以表达文本主题的关键词或词组是任务的关键。

针对上述所提问题,本文提出了一种以图神经网络为框架结合注意力机制的关键词生成模型——ADGCN,有效捕获关键词与文本主题间的关联密度,摆脱数据领域的偏向,以提高无监督学习方法的性能。基于图到序列的学习模型,在文本数据结构上应用图神经网络的同时结合注意力机制捕获数据中丰富的语义信息,自适应地将局部特征与全局依赖性相结合,使模型有效学习节点间的关联强度,更有效地利用图结构中的信息。

本文的主要贡献如下:

- 1) 构建文本内容与主题标签间的图链接,通过图构建层来表征文本的内部结构和主题间的关系。在图构建层的基础上采用注意力机制来学习节点特征空间的相关性,关注不同子空间的信息,引导节点间的关联密度。
- 2) 提出了一种新的线性解码模型,采用图神经网络编码图信息,捕获节点间的远依赖信息,解决未登录词的问题。
- 3) 收集并公布了一个社交平台的标签数据集。

1 相关工作

1.1 图神经网络

图神经网络模型在图形结构数据建模方面的应用越来越受到人们的关注,包括关系结构明确

的社会网络预测系统^[16]、推荐系统^[17-18]、知识图谱^[19]等,也包括关系结构不明确的非结构化场景,如图像分类^[20-21]、文本处理^[22-27]等。本文利用图神经网络对非结构化的文本数据进行建模。

近期研究致力于将图神经网络应用于文本分类任务中,涉及到将文本建模为图。文献[23]提出将文档转换为单词共现图,再将其用作图卷积层的输入,利用图表示的优势,有效捕获非连续和长距离语义的优势。文献[24]以概念图的形式表示文本,通过图卷积神经网络(graph convolution neural network, GCN)汇总匹配信号,比较包含相同概念顶点的句子来匹配一对文本。文献[25]构造了一个文本图张量,从文本的语义、句法和上下文信息 3 个角度来学习图内传播和图间传播,用于协调和集成图之间的异构信息。本文从以上工作得到启发,但是这些工作是为分类任务设计的,分类任务和生成任务是存在差别的。

一些研究是基于 GCN 做生成任务的。文献[26-27]提出了一种基于图序列的 Graph2Seq 神经网络模型,结合注意力机制并遵循传统的编码器方法,构建了一个图形编码器和一个序列编码器,基于图神经网络学习节点间属性在 SQL-to-text 任务中对 SQL 查询进行编码。文献[28]提出了基于图神经网络解决 AMR-to-Text 的问题。尽管这些工作将图神经网络做为编码器,但都是利用已经存在的图形式的信息,如 SQL 查询、抽象语义(abstract meaning representation, AMR)图和依存图,并且输入的文本相对社交平台发文较长,而笔者对短文本处理的工作更具挑战性。

1.2 关键词提取

关键词提取有利于信息检索、自动索引、自动分类、自动聚类、自动过滤等应用。文本中的主题不唯一,若只集中在文本语料库上,可能会导致知识获取的瓶颈和误解。目前的方法主要集中在怎样提取表达文本主题的关键词或词组。

文献[29]提出了一个利用主题模型从大众分类提取主题关键概念的方法。根据标签相对于某个主题的重要性进行排序,将排名靠前的标签作为提取主题关键的概念。文献[30]提出了一种以图表示文档、以节点表示文档单词、以边表示文档单词之间关系的关键字提取算法,用度和临近度中心测量主题关键词。文献[31]提出了一种利用 3 个主要因素不同于其他关键词提取方法,这 3 个主要因素是前面话题的时间、话题关联和参与者。当前发表言论的人应该被认为比其他参与者的言论更重要。本文从以上工作中得到启

发,将文本主题与图表示序列模型联合生成短文本关键词。

2 ADGCN 模型

2.1 模型概述

模型的目标是对句子进行信息压缩,输出短文本关键词信息。如图 1 所示,模型主要包括文本嵌入层、图构建层、注意力层、主题交互层和密集连接层,对应关键词生成的 3 个步骤:文本嵌入、特征编码和特征解码。

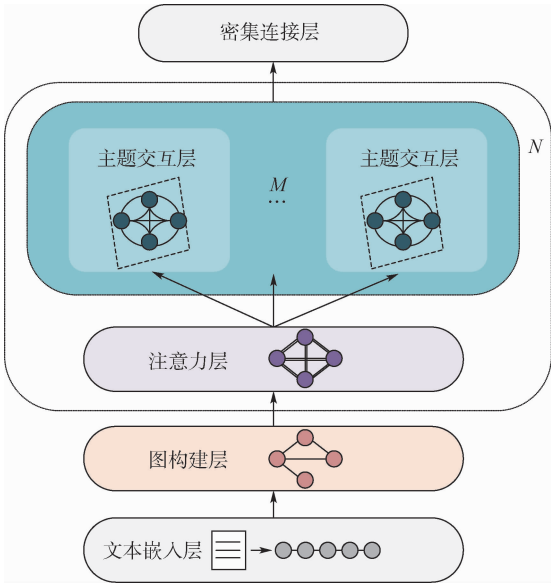


图 1 ADGCN 模型原理结构
Fig. 1 Schematic diagram of ADGCN model

文本嵌入部分的输入是预处理好的社交平台发文文本。对文本采用命名实体识别,并利用 TextRank 提取文本中的 M 关键词,对分好的词采用 glove 训练生成词嵌入向量。特征编码部分的目的是将文本嵌入层的输出进一步编码成一组隐藏向量,以获取文本节点关联信息和关键词隐藏

信息。图构建层对文本节点进行编码,用于组织文本句子间的关系。图构建后由 N 个相同的编码模块连接,每个块包含 2 个层:注意力层和主题交互层。注意力层对图加强边的权重,用于引导节点间的联系密度,这里采用多头注意力机制,挖掘隐含的节点关系,生成的是全连接的边权图。主题交互层对图结构进一步编码,编码框架采用 GCN,初始的输入是节点向量和节点的邻接矩阵,每一层包含上一层和初始时的输入。本文由 M 个主题交互层组成,对应着 M 头注意力。输出的隐藏向量由特征解码部分的密集连接层处理生成关键词。密集连接层对编码信息进行解码,在解码过程中采用注意力机制计算上下文向量对主题交互层的输出,并用一种复制机制来计算关键词的复制概率。

2.2 图构建层

图构建的目的是将短文本组织成图结构,便于模型理解文本的内部结构和主题之间的联系。相比较序列标注模型对文本的处理方式,图结构可以直接获取节点依赖关系,编码较广范围的语义信息,做到减少噪声语句对关键语句的影响,降低高维度特征的复杂度,缓解非主题数据对关键节点的扰动。

如图 2 所示,对含有 n 条句子的短文本 $S = \{s_1, s_2, \dots, s_n\}$ 构建图,节点分为关键节点 v_k 、空闲节点 v_{empty} 和主题节点 v_{title} 。关键节点是由关键词 $K = \{k_1, k_2, \dots, k_m\}$ 和包含关键词的句子组成。空闲节点 v_{empty} 是由剩余未包含关键词的句子组成。主题节点 v_{title} 是由 TextRank 计算一个文本主题 title,同时融合 HashTag 向量表征。若节点间包含相同的句子,则添加边,相同句子的个数作为边的权重。依据图构建算法生成图节点和边,并将图转化为邻接矩阵,作为下一层的输入。图构建方法如算法 1 所示。

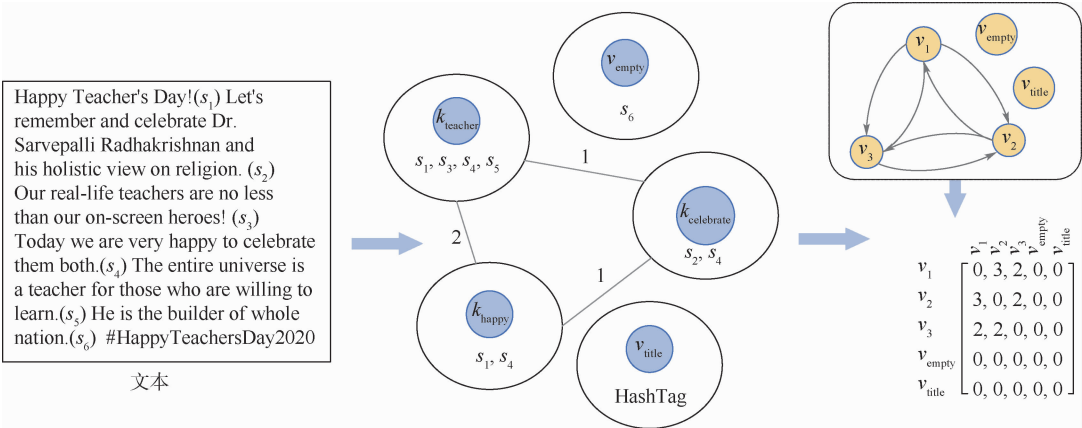


图 2 图构建过程
Fig. 2 Graph construction process

算法 1 图的构建方法。

输入: 关键词 $k = \{k_1, k_2, \dots, k_m\}$, 句子 $s = \{s_1, s_2, \dots, s_n\}$, 文本主题 title 。

输出: 关键节点 $\{v_1, v_2, \dots, v_m\}$, 空闲节点 v_{empty} , 主题节点 v_{title} , 节点 v_i 和节点 v_j 边的权重 w_{ij} 。

- 1: 对文本主题提取, 并对文本分词。
- 2: 做命名实体识别, 并提取文本中的 m 个关键词 $k = \{k_1, k_2, \dots, k_m\}$ 。
- 3: for $s = \{s_1, s_2, \dots, s_n\}$ do:
- 4: if 句子 s_i 包含关键词 k_i then:
- 5: 将 s_i 分配给节点 v_i
- 6: else:
- 7: 将 s_i 分配给节点 v_{empty}
- 8: end for
- 9: for 节点 v_i 和节点 v_j do:
- 10: w_{ij} 为 v_i 和 v_j 间包含相同句子的个数

11: end for

文本中词的初始输入是其在句子中的位置特征 $p_i \in \mathbf{R}^{d \times d}$ 及词特征 $x_i \in \mathbf{R}^{d \times d}$ 表征, 如下:

$$e_i = x_i + p_i \quad (1)$$

2.3 注意力层

注意力层用于引导节点间的关联密度。如图 3 所示, 对图构建层输出的邻接矩阵做多头注意力来关注不同子空间的信息, 生成全连接的注意力矩阵 $\tilde{\mathbf{A}}$, 对应邻接矩阵的边权连通图, 矩阵中每个元素对应相应节点之间边的权重, 从而捕捉到领域节点间的关系密度。

第 1 层的 t-head 注意力矩阵 $\tilde{\mathbf{A}}^n$ 的计算包括一个查询和一组键值对, 如下:

$$\tilde{\mathbf{A}}^n = \text{soft max} \left(\frac{Q\mathbf{W}_i^Q \times (K\mathbf{W}_i^K)^T}{\sqrt{d}} \right) \quad (2)$$

式中: Q 和 K 为模型第 1-1 层的表征; $\mathbf{W}^Q \in \mathbf{R}^{d \times d}$ 和 $\mathbf{W}^K \in \mathbf{R}^{d \times d}$ 为可学习参数矩阵。

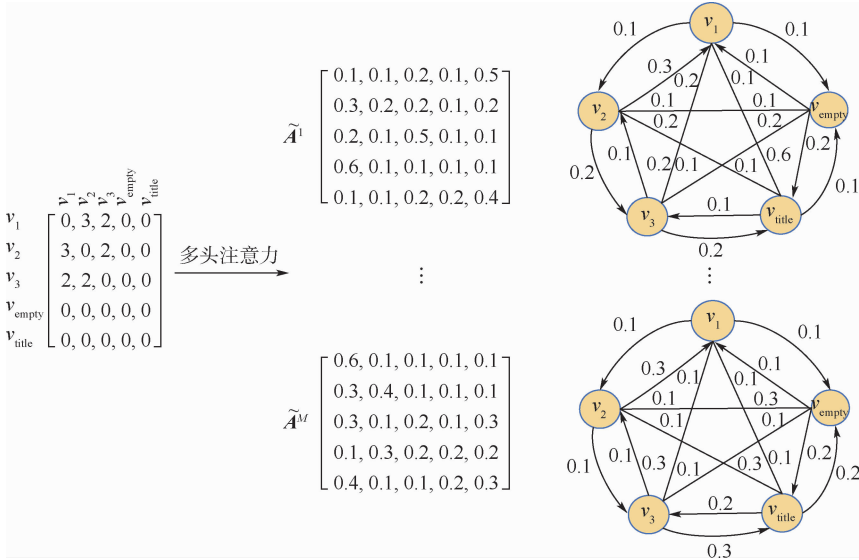


图 3 注意力层

Fig. 3 Attention layer

2.4 主题交互层

主题交互层是对图信息进行编码, 编码器采用 GCN^[32]。图神经网络既可以对节点内容编码, 又可以利用图结构中丰富的信息。其核心思想是: 通过学习一个映射函数映射图中的节点 v_i, v_j 可以聚合自己的特征与其邻居特征来生成节点 v_i 新的表示。通过多个 GCN 层可以挖掘到更深层次的节点信息, 节点的最终表征包含了更远的邻居节点信息。

在关键词生成任务中, 目的是预测出连贯的包含主题的词或短语 $k \in K$ (关键词集合)。对图中的 n 个节点, 可以表示成一个 $n \times n$ 的邻接矩阵 $\mathbf{A}$ 。 A_{ij} 和 A_{ji} 表示节点 i 和节点 j 之间边的权重值,

本文采用注意力层输出的 $\tilde{\mathbf{A}}$ 作为图嵌入的邻接矩阵。在 L 层的 GCN 中, 给定输入集合 $\mathbf{H}^0 = \{\mathbf{h}_1^0, \dots, \mathbf{h}_n^0\}$, 目的是输出 $\mathbf{H}^L = \{\mathbf{h}_1^L, \dots, \mathbf{h}_n^L\}$ 。第 1 层的节点 i 的输出向量 $\mathbf{h}_i^1$ 由第 1-1 层的节点 i 及其相邻节点表示, 如下:

$$\mathbf{h}_i^1 = \sigma \left(\sum_{j=1}^n \tilde{A}_{ij} \mathbf{W}^{(1)} \mathbf{h}_j^{l-1} + \mathbf{b}^{(1)} \right) \quad (3)$$

式中: σ 为激活函数; $\mathbf{W}^{(1)}$ 为做线性变换的权重矩阵; $\mathbf{b}^{(1)}$ 为偏差向量; $\mathbf{h}_i^0$ 为初始时输入的 $v_i, v_j \in \mathbf{R}^d$, d 为输入特征的维度。

随着 GCN 网络层数和迭代次数的增加, 节点的隐层表征会趋向于收敛至空间中的同一个位

置,即过度平滑问题。为了避免 GCN 过度平滑的问题,在每一层输入时考虑之前层的输出:

$$\mathbf{g}^l = [v_j; \mathbf{h}_j^1; \cdots; \mathbf{h}_j^{l-1}] \tag{4}$$

$$\mathcal{G}^{\text{out}} = \text{FFNN}(g_1, g_2, \cdots, g_n) \tag{5}$$

2.5 密集连接层

密集连接层完成对文本解码的工作,参考机器翻译模型^[33]采用循环神经网络(recurrent neural network, RNN)模型作为解码器。GCN 层的输出 $\mathcal{G}^{\text{out}}$ 作为密集连接层的初始输入,以生成一个注释标记序列 $\mathcal{Y} = \{y_1, y_2, \cdots, y_m\}$ 。考虑短文本主题是一个重要的信息,将图编码器处理后的主题隐藏特征作为解码器 RNN 的初始输入 t_0 , RNN 每一步的解码特征 t_i 。在解码过程中,采用注意力机制 α_j 计算上下文向量 $\mathbf{c}_i$ 来关注 GCN 的输出:

$$t_i = \text{RNN}(t_{i-1}, e_{i-1}) \tag{6}$$

$$\alpha_j = \frac{\exp(\delta(t_i, g_j))}{\sum \exp(\delta(t_i, g_k))} \tag{7}$$

$$\mathbf{c}_i = \sum \alpha_j \times g_j \tag{8}$$

式中: δ 为函数,计算当前解码器状态与原文各单词的相关度。

考虑初始抽取关键词本身的重要性,采用复制机制^[34],将预测的词标记的概率分布与注意力分布相结合,利用解码器的隐藏状态 t_i 和上下文信息向量 $\mathbf{c}_i$ 动态计算关键词的复制概率 p_{copy} :

$$y_i = \text{soft max}(U(\tanh(M([t_i; \mathbf{c}_i])) + b)) \tag{9}$$

$$p_{\text{copy}} = \sigma(W_{\text{copy}}[t_i; \mathbf{c}_i]) \tag{10}$$

$$p = (1 - p_{\text{copy}}) \times \mathcal{Y} + p_{\text{copy}} \times \alpha \tag{11}$$

式中: U 、 M 、 W_{copy} 和 b 为可学习的模型参数。

3 实验与分析

3.1 实验数据

本文从某个社交平台上收集数据,由于在不同主题下数据量差别很大,从生活、艺术、时尚、世界、科技、商业、微博、体育、健康、教育、旅行等主题中选择 3 个比较流行的主题:生活、教育和健康。对原文本处理成社交平台发文文本和话题标签的格式,称之为标签数据集 (TH)。TH 数据集详情如表 1 所示,可通过 <https://github.com/MinniEcho/keywords-generation.git> 访问该数据集,可

表 1 不同话题中的 TH 样本和标签数量

Table 1 Number of TH samples and tags in different topics		
主题	文本条数	标签个数
生活	14 044	10
教育	11 181	7
健康	4 868	5

以用来做提取标签或是文本分类等任务,这些数据是可以免费获取的。在数据集中存在一个话题共用多个主题的情况,如 #HappyTeachersDay 属于生活也属于教育。教育和健康 2 个主题的样本数相对生活的少一些,健康主题的样本数不足教育主题的样本数的一半,但是总样本数为 30 093 条,足以支撑模型训练的。

在收集的小型数据集上评估时,体现为:具有较好的可移植性和在大规模数据集中同样可以取得较好的结果,选择在一个公开数据集即 KP20k^[35]上进行测试和评估。该数据集是 Meng 等^[35]从 ACM、SicenceDirect 和 Web of Science 图书馆中收集的有关计算机领域的高质量数据,许多模型^[35-37]在该数据集上验证其模型性能的有效性。KP20k 数据集总样本数超过 50 万,是 TH 数据集的 18 倍,可以作为大规模数据集去验证模型性能。KP20k 详情如表 2 所示。

表 2 KP20k 数据集描述
Table 2 KP20k dataset description

KP20k	样本条数
训练集	530 809
验证集	20 000
测试集	20 000

3.2 实验设置

通过在验证集上的测试发现,从 $N = \{1, 2, 3\}$ 中选择编码部分的块数,从 $M = \{1, 2, 3, 4\}$ 选择多头注意力的头数,从 $L = \{2, 3, 4, 5\}$ 中选择主题交互层的层数。通过对验证集的初步实验,发现 ($N = 2, M = 2, L = 1, d = 300$) 的组合设置在 TH 数据集上取得了最好的效果,在顶点数相对较小的情况下,降低 L 数值,会缓解模型过度平滑问题。($N = 2, M = 3, L = 2, d = 300$) 的组合设置在 KP20k 数据集上取得了最好的效果。该模型是在 NVIDIA GeForce GTX 1650 下采用 CUDA9.0 训练。训练周期为 100,每 10 个周期用时约 270 s。词典大小为 60 000。句子大小为 100,生成关键词的长度最大值为 4。对于解码器,采用双向 LSTM,128 隐藏节点,2 层。dropout 率设置为 0.1,并采用 Adam 优化器训练参数。初始化的学习率为 0.01,epoch 为 10,每 epoch 一次学习率减少至原来的一半。

3.3 评估度量标准

本文采用 3 种度量方法评估收集的数据集上生成的关键词的质量,分别为相关性、信息量和连贯性。对所有的度量方法,让评分者按 3 种方式打分,分值范围为 0 ~ 10。在 KP20k 数据集上采用精准率 P 、召回率 R 和 F_1 值作为评估模型性能

的评估指标。

1) 相关性。用于评估结果是否与文本主题一致,即关键词是否符合社交平台发文文本内容,属于或是不属于社交平台发文文本。

2) 信息量。用于评估结果包含了多少具体信息。衡量的是输出结果是否涉及到某个人物或事件的特定方面,或者是对某个人物或事件的概括描述,或者是一般关键词。

3) 连贯性。用于评估输出的词是否流畅,是否是正确的、有意义的词。

本文邀请 3 位评分员去评估不同模型生成的关键词。由于评估过程复杂,选取 3 个话题的各 100 条发文的结果进行评估。评分员会得到原文、发文主题和关键词,这与用户在线阅读发文的方式相同。使用 Spearman 等级相关分数来衡量评分者之间的相关性, p -values 都低于 1×10^{-50} ,这表明评分员之间的评分有较好的相关性。在这些指标中,连贯性的差异更大。这和预期的一致,因为该指标更灵活,不同的人会有不同的评估标准。

3.4 Baseline

本文从无监督和有监督模型进行对比,选取的 Baseline 是 Tf-idf、TextTank、Maui^[36]、RNN^[35]、CopyRNN^[35]、CovRNN^[37]。Tf-idf、TextTank 属于无监督模型, Maui、RNN、CopyRNN、CovRNN 是有监督模型。Tf-idf、TextTank 是基于统计的方法,其针对候选关键词的次数来作为对文本的重要性,无法体现文本内容包含的语义信息,更无法生成未出现的关键短语。Maui、RNN、CopyRNN 和 CovRNN 是生成方法,可以捕捉文本上下文信息。

3.5 实验结果

表 3~表 5 展示了在生活、教育和健康 3 个不同主题下的不同 Baseline 模型与本文模型在不同度量下的结果。从结果中可以看出,本文提出的 ADGCN 模型在相关性和信息性方面都优于所有的 Baseline。

表 3 生活主题下 Baseline 模型和本文模型的 3 个度量评估比较

Table 3 Comparison of three measurement evaluation of Baseline model and proposed model under life topic				
模型	相关性	信息量	连贯性	结果
Tf-idf	6.75	5.70	8.03	6.83
TextTank	6.31	4.75	8.22	6.43
Maui	5.27	4.09	7.82	5.73
RNN	5.38	3.46	7.93	5.59
CopyRNN	6.52	5.21	8.04	6.59
CovRNN	6.56	5.24	8.09	6.63
ADGCN	8.13	6.21	7.53	7.29

表 4 教育主题下 Baseline 模型和本文模型的 3 个度量评估比较

Table 4 Comparison of three measurement evaluation of Baseline model and proposed model under education topic				
模型	相关性	信息量	连贯性	结果
Tf-idf	5.01	4.01	7.17	5.40
TextTank	4.93	4.47	7.34	5.58
Maui	4.31	4.39	5.35	4.68
RNN	4.05	4.60	5.67	4.77
CopyRNN	5.31	4.95	6.26	5.51
CovRNN	5.27	5.02	6.24	5.51
ADGCN	7.91	6.33	6.14	6.79

表 5 健康主题下 Baseline 模型和本文所提模型的 3 个度量评估比较

Table 5 Comparison of three measurement evaluation of Baseline model and proposed model under health topic				
模型	相关性	信息量	连贯性	结果
Tf-idf	4.65	4.51	6.95	5.37
TextTank	4.74	4.53	7.04	5.44
Maui	3.43	4.57	4.94	4.31
RNN	2.51	5.08	5.26	4.28
CopyRNN	4.87	5.39	5.31	5.19
CovRNN	4.85	5.47	5.43	5.25
ADGCN	6.97	6.36	5.09	6.14

1) 相关性。与所有其他 Baseline 模型相比,本文模型在相关性方面获得了较高的分数。这说明本文所提出的基于图到序列的学习模型可以更有效地获取和生成与文本主题相关的关键词。在遇到低频率话题时,如#HappeningNow 下的某个句子“The protesters in Rochester NY are ‘shutting down restaurants’, tables are broken, people running off sc...”。句子与话题的相关性不明显,这误导关键词的生成,这种现象严重影响了关键词相关性的分值。在 Baseline 模型中 Tf-idf 取得了较高的结果,这说明对与话题相关性不明显的句子提取关键词时,在样本数量足够的情况下,基于统计的方法可以提取出与文本主题相关的词。而在样本数量较少的情况下,如表 5 所示,统计的方法便不如生成模型的结果了。生成模型可以输出文本中不存在的词组,来更好地表达主题。

2) 信息量。从结果中可以看出,随着样本数的减少,关键词包含的信息量随之增大。与所有其他 Baseline 模型相比,本文模型在信息量方面获得了更高的分数。信息量和相关性是有一定关系的,相关性较高的模型,其信息量越丰富。但是当样本较少时也会出现不稳定的情况,像在健康主题下,RNN 模型信息量与相关性没有呈现正比

的关系。原因是:在健康主题下,样本中低频率话题较多,文本主题和话题不匹配会对模型造成一定的扰动。从结果中可以看出,基于图的模型可以利用文本结构,捕获文本情节,有稳定的健壮性。

3) 连贯性。基于统计的方法是从原文中直接复制词,连贯性只受到分词的影响。因此,结果中也显示基于统计的模型的连贯性分值要远高于生成模型。在检查了部分生成的关键词后,总结了以下几条影响流畅度的原因:①机器对人名不能较准确地识别,如赵立、坚定,而赵立坚是一个人名。②存在连续生成相同词的问题,严重影响了连贯性分值,本文猜测这和复制机制有关,这也是下一步待解决的问题。③模型生成的关键词是有悖知识逻辑的,如 Railway Speak Up(铁路大声的说)。这一问题也是比较难解决的问题之一。

表 6 展示了 Baseline 模型与本文模型在 KP20k 数据集上的 P 、 R 和 F_1 值的对比结果。发现各模型的召回率都比较低,这表明模型将很多正样本赋值了错误的标签,这和数据集样本的质量有关。另外,数据集中主题不明确的句子会对模型造成较大的扰动。KP20k 数据集中的标签大部分是文本中不存在的,因此,这对统计模型直接提取造成了重创。对生成模型来说,所生成的标签存在重复、次序混乱等现象而导致错误结果。但是本文模型 ADGCN 仍取得了较好的结果。

表 6 KP20k 数据集上 Baseline 模型和本文模型的精确率、召回率和 F_1 值评估比较

Table 6 Comparison of precision, recall and F_1 evaluation of Baseline model and proposed model on KP20k dataset			
模型	P	R	F_1
Tf-idf	0.413	0.052	0.093
TextTank	0.309	0.054	0.092
Maui	0.564	0.125	0.205
RNN	0.581	0.126	0.208
CopyRNN	0.652	0.213	0.321
CovRNN	0.683	0.220	0.333
ADGCN	0.735	0.327	0.453

3.6 实验分析

3.6.1 消融分析

对模型各模块进行消融研究,评估结果如表 7 所示。从结果中发现,编码模块的注意力层和主题交互层对模型的贡献最大。去除注意力层,模型 F_1 值降低至 0.413,减少了 0.04。这说明注意力层有效发挥了捕捉领域节点间关系密度的作用。去除主题交互层,模型 F_1 值降低至 0.406,

Table 7 Ablation of ADGCN model			
模型	P	R	F_1
ADGCN	0.735	0.327	0.453
去除图构建层	0.603	0.329	0.426
去除注意力层	0.667	0.299	0.413
去除主题交互层	0.661	0.293	0.406
去除注意力层,主题交互层	0.565	0.301	0.393
去除密集连接层	0.682	0.326	0.441

减少了 0.047。这说明编码远距离的邻居节点信息可以提升模型准确率。若两者都去除, F_1 值则会降低 0.06,这对模型性能的影响是最大的。而图构建层和密集连接层对模型的影响相对较小。可见,注意力与图神经网络相结合有助于 ADGCN 学习短文本内的信息聚合,以生成更好的节点信息特征。

3.6.2 数据集规模对模型性能的影响

对 KP20k 数据集进行统一划分,设置 5 种规模训练数据:20%、40%、60%、80% 和 100% 的训练数据。对 ADGCN 和 Baseline 模型在不同训练设置上进行评估。如图 4 所示,在数据规模不大时,CovRNN 取得了较好的效果,这说明 CovRNN 可以从样本较少的数据中捕获有利的信息,但是与其他生成模型的差距并不明显。随着数据集规模的扩增,发现生成模型与统计模型的性能差距越明显,在数据集规模达到 40% 时,生成模型远超前于统计模型。在数据集规模达到 60% 时,本文所提出的 ADGCN 的 F_1 值达到了其他 Baseline 模型的最优值。这说明 ADGCN 可以更有效地利用训练数据集,在数据规模一定的情况下,基于图到序列的模型可以有效学习短文本中主题信息。

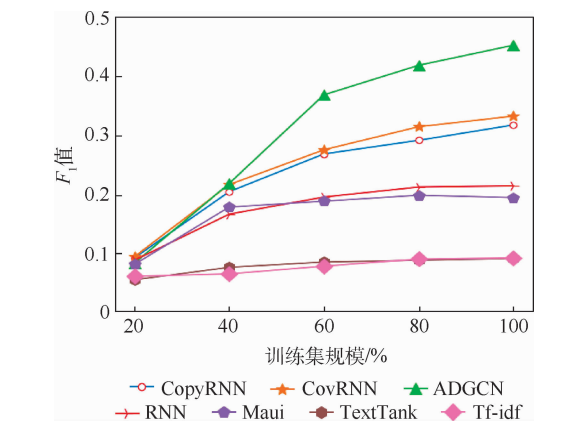


图 4 不同数据规模下各模型的 F_1 值比较
Fig. 4 Comparison of F_1 values of different models under different data scales

3.6.3 话题种类对模型性能的影响

评估 TH 数据集中话题种类对模型性能的影响。设置包含不同话题种类个数 1、2、3、4 和 ≥ 5 的句子,对 ADGCN 和与 ADGCN 较相近的 CopyRNN、CovRNN 模型在不同数量的话题句子上进行关键词质量评估,分别从关键词的相关性、信息量、连贯性 3 个方面进行统计评估。统计包含不同个数的话题标签的发文比例,如图 5 所示。话题种类个数为 1 的只占据了 60%,而话题种类个数在 5 之上的占据了 5%,这些数据增加了文本主题的噪声,对短文本主题表征会有一定的影响。如图 6 所示,发现各模型在话题种类个数 ≤ 2 时生成的关键词质量是最好的。而随着话题种类个数的增加都有减少的趋势,这说明话题种类的数量越多对模型的扰动性越大。

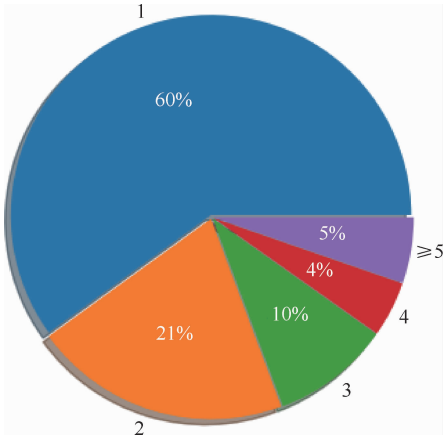


图 5 话题标签个数比例
Fig.5 Ratio of the number of topic tag

在对相关性评估分析时发现,ADGCN 的相关性分值虽然也会收到数据扰动的影响,但是相对 CovRNN 和 CopyRNN 两个模型较缓慢。CovRNN 和 CopyRNN 在话题标签个数大于 2 时急速下降,在[2~4]区间存在不明显的差距。而在对信息量评估时,2 个模型的差距明显化,这说明 CovRNN 相对 CopyRNN 是稳定的。当话题种类个数为 1 或是超过 5 时,无论是相关性还是信息量 CopyRNN 的表现要比 CovRNN 模型弱一些,猜测可能和数据不平衡有关。而 ADGCN 可以有效缓解数据不平衡的影响,因此表现是最优的。

在对连贯性评估分析时发现,CovRNN 和 CopyRNN 在话题种类个数为 1 时的表现比 ADGCN 好,但是当话题标签个数增加时,2 个模型下降明显,而 ADGCN 下降缓慢,并在数值较大时有缓升的趋势。这说明在文本组织结构明显的情况下,序列标注模型对文本信息的组织可能要优于图结构模型。对结构化不明显的文本,图结

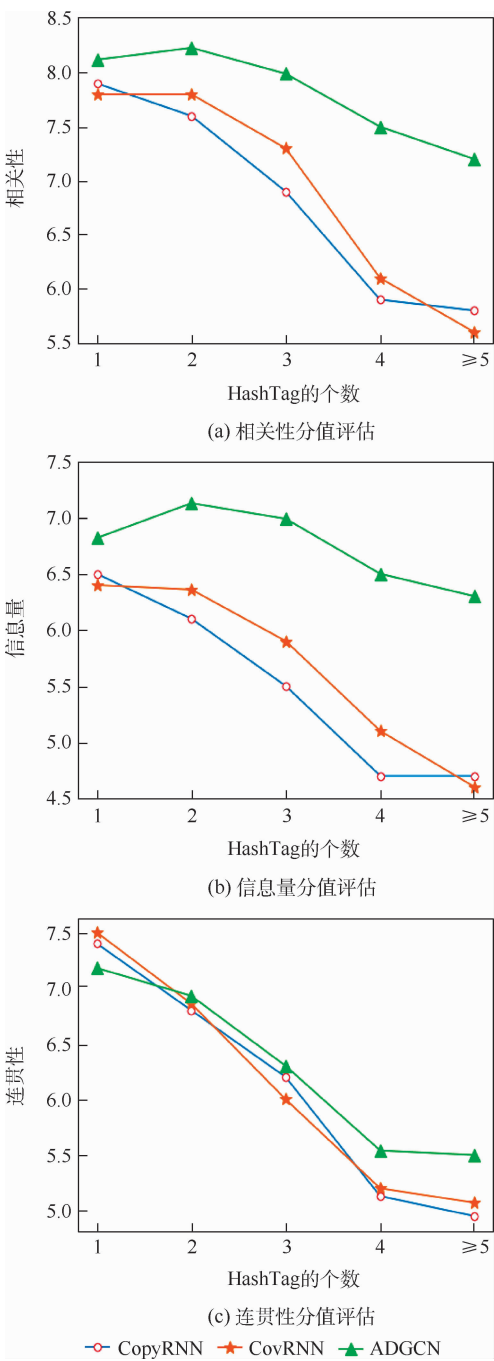


图 6 不同话题标签个数下度量值的评估
Fig.6 Evaluation of measurement value under different numbers of topic tag

构模型表现出了优势。

造成数据扰动的原因可能有以下 2 点:①因发文有字符字数限制,话题种类数量越多,文本的组织结构会越差。这对序列标注模型 RNN 的影响是较大的。相比较,利用图表示模型的结果会好一点。②随着话题种类的增加,主题节点表征的信息特征越不明显,这大大降低了文本主题的表达,模型信息获取的性能会随之下降。而本文所提出的 ADGCN 在不同设置上的性能完全优于其他生成模型。

4 结 束 语

本文提出了一种新的图到序列的关键词生成,一是通过图构建层来表征文本内容与主题标签间的图链接;二是结合注意力机制来学习节点特征空间的相关性,关注不同子空间的信息,引导节点间的关联密度。该方法有效缓解了短文本主题依赖和文本组织结构差对模型造成扰动的影响。ADGCN 模型不仅具有良好的相关性、信息量及连贯性,在大规模数据集 KP20k 上也取得了较好的性能。

接下来会对模型的不足进行改进和优化,如复制机制的影响、知识逻辑的推理等。缓解多话题标签对文本主题的影响也是未来待解决的问题之一。

参考文献 (References)

[1] BOUDIN F. A comparison of centrality measures for graph-based keyphrase extraction[C] // Proceedings of the International Joint Conferences on Natural Language Processing (IJCNLP), 2013: 834-838.

[2] LAHIRI S, CHOUDHURY S R, CARAGEA C. Keyword and keyphrase extraction using centrality measures on collocation networks[EB/OL]. (2014-01-25) [2020-10-01]. <https://arxiv.org/abs/1401.6571>.

[3] PALSHIKAR G K. Keyword extraction from a single document using centrality measures[J]. Pattern Recognition and Machine Intelligence (PReMI), 2007: 4851 (1): 503-510.

[4] EDIGER D, JIANG E J, BADER D A, et al. Massive social network analysis: Mining twitter for social good[C] // 39th International Conference on Parallel Processing (ICPP), 2010: 583-593.

[5] BULGAROV F, CARAGEA C. A comparison of supervised keyphrase extraction models[C] // Proceedings of the 2015 International Conference on World Wide Web, 2015: 13-14.

[6] MOTHE J, RAMIANDRISOA F, RASOLOMANANA M. Automatic keyphrase extraction using graph-based methods[C] // Proceedings of the 33rd Annual ACM Symposium on Applied Computing, 2018: 728-730.

[7] 刘啸剑, 谢飞, 吴信东. 基于图和 LDA 主题模型的关键词抽取算法[J]. 情报学报, 2016, 35 (6): 664-672.

LIU X J, XIE F, WU X D. Keyword extraction algorithm based on graph and LDA topic model[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35 (6): 664-672 (in Chinese).

[8] BOUDIN F. Unsupervised keyphrase extraction with multipartite graphs[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 667-672.

[9] SAROJ K B, MONALI B, JACOB S. A graph based keyword extraction model using collective node weight[J]. Expert Systems with Application, 2018, 97 (1): 51-59.

[10] BELLAACHIA A, AL-DHELAAN M. NE-Rank: A novel graph-based key phrase extraction in twitter[C] // Proceedings of the International Joint Conferences on Web Intelligence and Intelligent Agent Technology, 2012: 372-379.

[11] LI Z, WANG C. Keyword extraction with character-level convolutional neural tensor networks[C] // 23rd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2019: 400-413.

[12] 杨丹浩, 吴岳辛, 范春晓. 一种基于注意力机制的中文短文本关键词提取模型[J]. 计算机科学, 2020, 41 (1): 193-198.

YANG D H, WU Y X, FAN C X. A Chinese short text keyword extraction model based on attention mechanism[J]. Computer Science, 2020, 41 (1): 193-198 (in Chinese).

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017: 5998-6008.

[14] 冯建周, 宋沙沙, 王元卓, 等. 基于改进注意力机制的实体关系抽取方法[J]. 电子学报, 2019, 47 (8): 1692-1700.

FENG J Z, SONG S S, WANG Y Z, et al. Entity relation extraction method based on improved attention mechanism[J]. Acta Electronica Sinica, 2019, 47 (8): 1692-1700 (in Chinese).

[15] MATTHEW E P, MARK N, MOHIT I, et al. Deep contextualized word representations[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 2227-2237.

[16] HAMILTON W L, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C] // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017: 1024-1034.

[17] BERG R, KIPF T N, WELLING M. Graph convolutional matrix completion[EB/OL]. (2017-10-25) [2020-10-01]. <https://arxiv.org/abs/1706.02263>.

[18] YING R, HE R, CHEN K, et al. Graph convolutional neural networks for web-scale recommender systems[C] // Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining, 2018: 974-983.

[19] HAMAGUCHI T, OIWA H, SHIMBO M, et al. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach[C] // Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017: 1802-1808.

[20] KAMPFFMEYER M, CHEN Y, LIANG X, et al. Rethinking knowledge graph propagation for zero-shot learning[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 11487-11496.

[21] WANG X L, YE Y F, GUPTA A. Zero-shot recognition via semantic embeddings and knowledge graphs[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6857-6866.

[22] LI Z, SUN Y, ZHU J, et al. Improve relation extraction with dual attention-guided graph convolutional networks[J]. Neural Computing and Applications, 2021, 33: 1773-1784.

[23] PENG H, LI J X, HE Y, et al. Large-scale hierarchical text classification with recursively regularized deep graph-CNN[C] //

Proceedings of the 2018 World Wide Web Conference, 2018: 1063-1072.

[24] LIU B, NIU D, WEI H, et al. Matching article pairs with graphical decomposition and convolutions [C] // Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019: 6284-6294.

[25] LIU X, YOU X, ZHANG X, et al. Tensor graph convolutional networks for text classification [C] // The Thirty-Second Innovative Applications of Artificial Intelligence Conference, 2020: 8409-8416.

[26] XU K, WU L F, WANG Z G, et al. Graph2Seq: Graph to sequence learning with attention-based neural networks [EB/OL]. (2018-12-03) [2020-10-01]. <https://arxiv.org/abs/1804.00823>.

[27] XU K, WU L, WANG Z, et al. SQL-to-text generation with graph-to-sequence model [C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 931-936.

[28] BECK D, HAFFARI G, COHN T. Graph-to-sequence learning using gated graph neural networks [C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 273-283.

[29] XUE H, QIN B, LIU T. Topical key concept extraction from folksonomy through graph-based ranking [J]. Multimedia Tools and Applications, 2016, 75 (15) : 8875-8893.

[30] NAGARAJAN R, NAIR S A H, ARUNA P. Keyword extraction using graph based approach [J]. International Journal Advanced Research in Computer Science and Software Engineering, 2016, 6 (10) : 25-29.

[31] SONG H, GO J, PARK S, et al. A just-in-time keyword extraction from meeting transcripts using temporal and participant information [J]. Journal of Intelligent Information Systems, 2017, 48 (1) : 117-140.

[32] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [EB/OL]. (2017-02-22) [2020-10-01]. <https://arxiv.org/abs/1609.02907>.

[33] DZMITRY B, KYUNGHYUN C, YOSHUA B. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2020-10-01]. <https://arxiv.org/abs/1409.0473>.

[34] GU J T, LU Z D, LI H, et al. Incorporating copying mechanism in sequence-to-sequence learning [EB/OL]. (2016-06-08) [2020-10-01]. <https://arxiv.org/abs/1603.06393>.

[35] MENG R, ZHAO S, HAN S, et al. Deep keyphrase generation [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2017: 582-592.

[36] MEDELYAN O, FRANK E, WITTEN I H. Human-competitive tagging using automatic keyphrase extraction [C] // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 1318-1327.

[37] ZHANG Y, XIAO W. Keyphrase generation based on deep Seq2seq model [J]. IEEE Access, 2018, 6: 46047-46057.

Adaptive short text keyword generation model

WANG Yongjian, SUN Yaru, YANG Ying*

(The Third Research Institute of Ministry of Public Security, Shanghai 201204, China)

Abstract: Keyword extraction has a great impact on text processing, and the accuracy and fluency of keyword recognition are the keys to the task. In order to effectively solve the problems such as inaccurate word division, mismatch between keywords and text topics, and multi-language mixing in the process of keyword extraction from short text, we propose an adaptive short text keyword generation model based on graph convolutional neural network (ADGCN). First, the model uses graph neural network as the coding framework of text information feature extraction to solve the problem of irregular short text structure and the existence of complex information between words. Then, according to the location features and context features of words, the self attention mechanism is combined to capture rich context dependent information. Finally, a linear decoding scheme is used to generate interpretable keywords. We collect and publish a tag dataset TH from social media platform, including text and topic tags. We evaluate and analyze the relevance, information and coherence of the model results from the perspective of user needs. The model can not only generate keywords that meet the topic of short text, but also effectively alleviate the impact of data disturbance on the model. It is proved that the model performs well on the public dataset KP20k and has good portability.

Keywords: keyword extraction; keyword generation; graph neural network; attention mechanism; topic model

http://bxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0538

基于变分推理的网络舆情传播模式分类

唐红梅^{1,2}, 唐文忠¹, 李瑞晨¹, 王衍洋^{3,4,*}, 王丽宏⁵

(1. 北京航空航天大学 计算机学院, 北京 100083; 2. 新疆维吾尔自治区科技项目服务中心, 乌鲁木齐 830000;
3. 北京航空航天大学 航空科学与工程学院, 北京 100083; 4. 北航江西研究院, 南昌 330096;
5. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 随着网络社交媒体的快速发展,对舆情信息的传播模式进行分析成为研究热点。针对网络舆情传播模式分类任务中,小样本数据多路径生成分类正确率低的问题,提出了舆情传播领域知识图谱结构定义,建立了基于微博数据的舆情传播知识图谱与舆情传播分析任务数据集,使用 GraphDIVA 模型进行舆情传播模式分类,并在自建数据集中进行了舆情传播模式分类 25 样本测试实验。结果表明:模型在经过 20 轮训练后,分类正确率从 76% 提升到 89.4%,说明 GraphDIVA 模型在减少训练次数、提升分类正确率方面具有更优的效果。

关 键 词: 舆情传播模式; 知识图谱; 知识图谱推理; 图神经网络; 模式分析

中图分类号: TP391

文献标志码: A **文章编号:** 1001-5965(2022)02-0209-08

当前,随着网络信息量的激增,信息生态环境受到污染,公众获取、辨别和消化信息的能力较弱,容易受到不实信息的影响,产生错误的情感导向^[1]。传统的舆情传播模式分析,主要集中在数学建模与历史数据挖掘,且多数基于统计学原理,缺少完整的舆情分析功能。随着知识图谱的发展,很多领域均引入知识图谱^[2]。

图模型对舆情的传播过程进行建模,将舆情信息的扩散以节点和边的形式进行表示,能完整建模舆情事件在用户间的传播扩散过程,并根据需要动态添加用户间关注关系、舆情信息引用关系等信息,提升分析的效果。王晰巍等^[3]基于社会网络分析法对舆情的传播过程进行建模,基于建立好的社会网络图,分析节点的入度与出度和节点的聚类情况,可以识别出传播过程中的关键用户。崔树娟等^[4]基于多子网复合复杂网络模型构建多关系社交网络,将复杂系统中个体间的

多种相互关系映射为向量空间的多维向量,并将复杂网络复合及分解转化为向量空间的基变换,从而为具有多种关系的复杂网络研究提供了新的解决方法。王兰成和娄国哲^[5]在 2018 年基于知识图谱对涉军舆情进行了分析,提出了网络舆情知识图谱的舆情事件表达和处理引擎设计。马哲坤和涂艳^[6]在 2019 年基于知识图谱进行了突发话题监测研究,发现基于知识图谱的话题监测相比于传统方法在准确率与召回率指标上有较大提升。基于知识图谱的舆情分析过程,就是舆情知识图谱的推理过程。Chen 等^[7]在 2018 年针对知识图谱链接预测提出的 DIVA 推理模型可以应用在舆情传播模式分类上,通过变分推理框架,综合之前基于路径推理与基于特征推理的优点,达到了更优的性能。受变分自编码器(VAE)^[8](提出于 2013 年)的启发,DIVA 模型在设计中使用变分自编码器结构完成了路径搜索与路径推理的

收稿日期: 2020-09-22; 录用日期: 2020-10-23; 网络出版时间: 2020-11-26 17:32
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201125.1441.001.html
基金项目: 新疆维吾尔自治区自然科学基金(2020D01A95)
* 通信作者。E-mail: wangyanyang@buaa.edu.cn

结合。

通过对已有模型进行研究对比,发现在舆情事件传播模式分类任务中,由于数据规模庞大,生成的舆情事件图谱包含上百万的实体,且实体间关系非常复杂,对推理模型有极高的要求。

在公开数据集的测试中,DIVA 模型针对较为简单的单一关系分类任务,只能对给定的 2 个实体在知识图谱中判断两者是否属于一种关系,并分别对属于与不属于 2 种情况给出概率。由于单独的路径搜索过程无法有效区分不同路径的质量高低,导致在复杂网络中难以找到高质量的连接路径,影响最终的识别正确率。因此,如何通过二分类模式由一个模型直接输出 2 种分类各自的概率,提升分类任务的正确率是本文的研究重点。

本文采用 GraphDIVA 模型进行舆情传播模式分类研究,从而降低了训练次数,提升了分类任务的准确性。

1 舆情传播知识图谱

1.1 GraphDIVA 模型

GraphDIVA 模型以变分推理 DIVA 模型为基础,从网络模型结构和计算过程 2 个维度进行分析,在路径推理模块中引入 GraphSAGE 算法,用于解决原模型缺少对于图结构特征的感知能力导致正确率下滑这一问题。GraphDIVA 模型在生成路径特征矩阵时,引入 GraphSAGE 算法,对路径所在的子图结构进行采样并聚合,过程如图 1 所示。

GraphSAGE 算法是 Hamilton 等^[9]在 2017 年提出的对于图结构特征进行感知的算法,对于动态图结构拥有非常好的适应性。GraphSAGE 算法认为一个节点的特征决定于周围节点,通过多轮采样聚合后便能得到最终的节点特征。同时,将特征的聚合操作定义为先从中心节点逐层向外扩散采样,再从最外层逐层向内聚合的递归过程,该方式可以解决图卷积神经网络(GCN)^[10]弹性不足的问题。

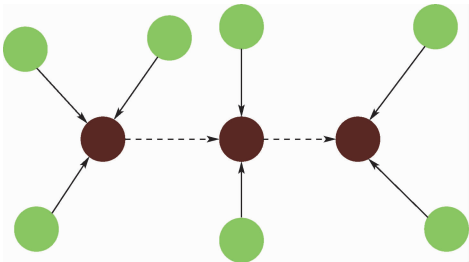


图 1 GraphDIVA 路径特征生成过程

Fig. 1 GraphDIVA path feature generation process

在使用卷积神经网络(CNN)窗口提取特征之前,先使用 GraphSAGE 算法对路径中的每个实体随机采样固定数据的数目与连接关系,再使用 MaxPooling 策略对采样到的嵌入向量进行聚合,得到路径实体与关系的特征表示。详细计算过程如下:

AGGREGATE(v) = $\max(\{\sigma(W_{pool}[r_{ui};e_{ui}] + b_{pool}), \forall u_i \in N(v)\})$ (1)

$f(v) = W_{neighbor} AGGREGATE(v) + W_{self}[r_v;e_v] + b_{pool}$ (2)

式中: $[r_v;e_v]$ 表示路径中某一步的关系嵌入向量与实体嵌入向量; $N(v)$ 为从路径实体出发随机选择多个邻接实体; $[r_{u_i};e_{u_i}]$ 表示与路径实体邻接的实体与连接关系嵌入向量; $f(v)$ 为最终输出的针对路径中某一步的特征向量。

最终产生的路径特征矩阵由路径中的多步特征向量构成。

GraphDIVA 路径推理模块结构如图 2 所示。

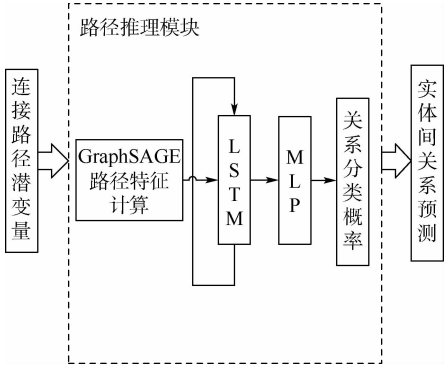


图 2 GraphDIVA 路径推理模块结构

Fig. 2 GraphDIVA path reasoning module structure

图 2 中:GraphSAGE 路径特征计算过程与 LSTM 网络^[11]为相对于原 DIVA 模型结构的变化。相比于原模型中使用的 3 个 CNN 窗口,GraphSAGE 算法使用 LSTM 计算路径的次序特征,最终将 LSTM 的输出向量作为路径的特征向量,如下所示:

$h_{i+1} = \text{LSTM}(h_i, (f(v_i))) \quad i \in 1, 2, \dots, n - 1$ (3)

$\text{feature} = \text{LSTM}(h_n, (f(v_n)))$ (4)

式中: h 为 LSTM 计算路径的每个时间步特征; v 为图中的一个节点。

相比于原 DIVA 模型直接将输入路径中实体与关系的嵌入向量作为路径特征矩阵的做法,调整后的 GraphDIVA 模型先通过 GraphSAGE 算法识别路径中实体所在子图的结构特征,再利用 LSTM 提取路径的次序特征,大大提升了模型对图的感知范围与能力,从而解决了原模型中存在

的正确率下滑等问题。

1.2 舆情图谱结构

与通俗意义上只注重建模领域概念的知识图谱不同,舆情知识图谱需要额外添加舆情的传播过程信息。娄国哲和王兰成^[12]对基于知识图谱的网络舆情建模方法进行了研究,认为面向网络舆情的知识图谱中实体分为事件标签、事件主体、事件客体与事件传播方 4 类。事件标签为对事件本身的定性描述与分类;事件主体为事件内容涉及到的单位与个人,如机构、企业、明星等;事件相关客体描述事件的客观信息,如发生的地点、涉及到的时间等;事件传播方为舆情在传播过程中涉及到的用户、站点等。同时,知识图谱中的关系分为事件发起方与事件的关系、事件间的关系与事件传播关系。其中事件发起方与事件的关系包括发起、关注、参与等;事件间关系包括引发、有关、相似等;事件传播关系包括发表、评论、转发、点赞等。

本文主要研究舆情在微博用户间的传播过程,不关注舆情事件的概念、涉及到的主体与时间地理因素,因此本文定义的舆情传播知识图谱的实体与关系在上述模型的基础上进行了大幅精简与调整。实体包含关键词、博文、评论、用户与话题 5 类,关系分为包含关键词、转发、评论、用户创建、提及话题 5 类。

对于用户的转发行为,舆情传播知识图谱的建模三元组如下所示:

$$(\text{user}_i, \text{create}, \text{blog}_a) \tag{5}$$

$$(\text{user}_j, \text{create}, \text{blog}_b) \tag{6}$$

$$(\text{blog}_a, \text{repost}, \text{blog}_b) \tag{7}$$

式中: user 表示用户; blog 表示博文。

首先,需要建立 2 个由用户指向各自创建的博文的三元组,然后,建立由被转发博文指向转发博文的三元组。

对于用户的评论行为,舆情传播知识图谱的建模三元组如下所示:

$$(\text{user}_j, \text{create}, \text{comment}_b) \tag{8}$$

$$(\text{blog}_a, \text{comment}, \text{comment}_b) \tag{9}$$

式中: comment 表示用户评论。

与转发行为类似,首先,需要建立从用户指向创建的评论的三元组,然后,建立由被评论博文指向评论的三元组。

对于博文与关键词的对应关系,舆情传播知识图谱的建模三元组如下所示:

$$(\text{keyword}_x, \text{keyword}, \text{blog}_a) \tag{10}$$

式中: keyword 表示博文的关键字。

对于博文与话题标签的对应关系,舆情传播知识图谱的建模三元组如下所示:

$$(\text{blog}_a, \text{reference}, \text{tag}_x) \tag{11}$$

式中: tag 表示博文的标签。

在进行舆情事件的划分时,直接使用关键词作为划分标准。如果从不同的实体出发,最终能到达相同的关键词实体,则认为这些实体都属于同一个舆情事件。

1.3 数据采集与处理

本文选择新浪微博官方热搜排行榜中热门的新闻类与娱乐类话题,提取核心关键词作为后续搜索内容的基础,每隔 3 ~ 5 天,根据热搜排行榜的变化,更换核心关键词,扩大舆情的搜索范围,获得最原始的舆情数据。

确定关键词后,以微博搜索作为入口,通过关键词对有关微博内容进行搜索,并按照热度提取排名前 10 的结果。对于每一篇微博,按热度获取热门评论与热门转发,再根据热门转发链接完成转发关系的递归获取操作。同时,在解析页面时,记录当前页面与上一个页面的链接,从而方便从结果中重建正确的微博转发关系。

本文研究数据来源于微博,时间段为 2019 年 9 月 28 日至 2019 年 11 月 4 日。数据结果类型包括数据数量、博文转发量、博文传播地位、博文话题标签、评论点赞量、话题标签数据、关键词出现次数和舆情传播深度等,详细数据结果如表 1 所示。博文数据的分布状况如表 2 ~ 表 4 所示。

表 1 数据数量统计

Table 1 Data quantity statistics	
类型	数量/条
博文	712 577
评论	388 689
用户	709 603
话题标签	32 549
关键词	74

表 2 博文转发量统计

Table 2 Blog post forwarding quantity statistics	
类型	数量/条
转发量为 0	651 502
转发量大于 0 小于 1 000	56 547
转发量大于或等于 1 000, 小于 10 000	3 275
转发量大于或等于 10 000, 小于 50 000	1 006
转发量大于或等于 50 000, 小于 200 000	201
转发量大于或等于 200 000	46
最大转发量	1 304 609

评论数据的分布状况如表 5 所示。话题标签数据的分布状况如表 6 所示。关键词数据的分布状况如表 7 所示。舆情传播深度与广度分布如表 8 所示。

表 3 博文传播地位统计

Table 3 Statistics of propagation status of blog posts

类型	数量/条
始发微博	98 173
转发微博	614 404

表 4 博文话题标签统计

Table 4 Blog post hashtag statistics

类型	数量/条
含话题标签的微博数目	96 209
平均话题标签数量	1. 66
最大话题标签数量	24

表 5 评论点赞量统计

Table 5 Statistics of the amount of likes

类型	数量/条
点赞量为 0	197 794
点赞量大于 0 小于 1 000	183 092
点赞量大于或等于 1 000, 小于 5 000	6 108
点赞量大于或等于 5 000, 小于 20 000	1 438
点赞量大于或等于 20 000, 小于 100 000	252
点赞量大于或等于 100 000	5
最大点赞量	416 050

表 6 话题标签数据统计

Table 6 Hashtag data statistics

类型	数量/条
出现次数为 1	23 142
出现次数大于 1, 小于 20	8 424
出现次数大于或等于 20, 小于 50	580
出现次数大于或等于 50, 小于 200	301
出现次数大于或等于 200, 小于 1 000	66
出现次数大于或等于 1 000	11
最大出现次数	5 863

表 7 关键词出现次数统计

Table 7 Statistics of keyword occurrence times

类型	数量/条
出现次数大于 1, 小于 100	4
出现次数大于或等于 100, 小于 400	23
出现次数大于或等于 400, 小于 1 000	23
出现次数大于或等于 1 000, 小于 4 000	18
出现次数大于或等于 4 000	6
最大出现次数	17 845

通过对以上统计数据进行分析可以看出,本文建立的舆情传播图谱覆盖范围广泛,不仅包含丰富的热门话题种类,而且还有效保留了舆情的原始传播过程,可以充分支撑舆情传播模式分类研究工作。

表 8 舆情传播深度统计

Table 8 Statistics of public opinion propagation depth

类型	数量/条	总转发数量/条
无转发博文数量	66 899	0
转发深度为 1	27 267	343 720
转发深度为 2	2 473	150 780
转发深度为 3	540	74 564
转发深度为 4	141	28 389
转发深度为 5	71	21 148
转发深度为 6	33	11 698
转发深度为 7	17	6 040
转发深度为 8	4	925
转发深度为 9	8	6 127
转发深度为 10	3	2 113
转发深度为 21	1	154

2 基于变分推理的传播模式分析

2.1 传播模式

对于舆情在用户间的传播特性,领域中已有大量的研究成果。刘继和李磊^[13]在 2013 年就针对微博的舆情传播过程对用户间的传播模式特点进行了分析。本文将舆情的传播关系网划分为单关键点型、多关键点型与链式型 3 种基本模式,结构示例如图 3 所示。

单关键点由一个中心用户与大量粉丝用户组成。中心用户往往在小范围社区内有非常高的人气,多为某一特定领域的“大 V”;粉丝用户对于中心用户有极高的信任与热情,大多会在第一时间转发相关舆情内容。在这种模式中,中心用户对于舆情事件的爆发有较强的控制力,能把握住舆情的走向节奏,但是最终的传播规模也往往只局限在某个小圈子里,缺少大规模传播的潜力。

多关键点由交换用户、多个中心用户与大量不同领域的普通用户组成,相当于多个单关键点的复合结构。交换用户为多关键点型的核心节点,类似于沟通不同领域的中央交换机,拥有与大量不同领域中心用户的紧密关系,能够发动多个

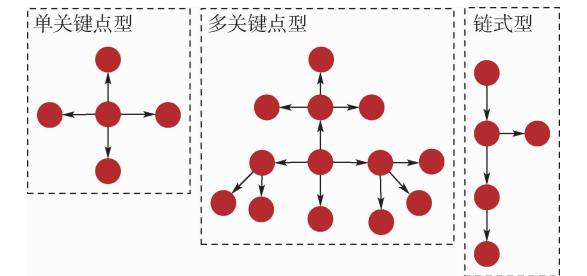


图 3 微博舆情信息传播基本模式示例

Fig.3 Example of basic pattern of Weibo public opinion information propagation

“大 V”传播自己的观点;与此同时,多个中心用户可以通过交换用户实现信息互动,增强对不同领域的传播能力。由于交换用户的存在,多关键点在传播时相比于单关键点型的覆盖范围大大延伸,往往能形成影响非常巨大的舆情事件。

链式与单关键点及多关键点在构成上有较大的差别。单关键点与多关键点总体上属于中央式结构,能将节点分为主从 2 种;而链式则更类似分布式结构,各个节点的地位大体是均等的。链式结构往往出现在有较强针对性的舆情传播过程中,波及到的用户也常常属于相近的社交圈。由于链式传播的路径与生命周期较长,容易造成舆情的演化与扭转,并引发新的舆情传播热点。

实际舆情事件中,以上 3 种传播模式不会以单独的形式存在,而是产生大量的混合型结构。一种常见的情况是先由普通网民发起链式传播,之后被“大 V”发现后转变为多关键点模式。

2.2 任务定义

知识图谱推理任务是指基于三元组进行推理的模型。对于任意一个知识图谱,包含实体集合 $\mathcal{E}$ 与关系集合 R 。知识图谱的结构则可以 (e_s, r, e_d) 的三元组进行表示。以舆情传播知识图谱为例,用户 A 与用户 B 表示为实体 $e_{\text{user_a}}$ 和 $e_{\text{user_b}}$,用户 B 转发用户 A 消息的操作表示为关系 r_{repost} ,舆情的扩散则可用三元组描述为 $(e_{\text{user_a}}, r_{\text{repost}}, e_{\text{user_b}})$ 。推理任务基于部分残缺的三元组,推测可能的答案。推理任务的三元组可能缺少关系 $(e_s, ?, e_d)$,也可能缺少实体 $(e_s, r, ?)$,对应了链接预测与实体预测 2 种子任务。本文基于舆情传播模式分析的场景,更关注于链接预测^[14],也就是推测用户 A 与用户 B 间可能存在的联系,作为舆情传播模式的划分依据。原始的 DIVA 模型针对较为简单的单一关系分类任务,只能对给定的 2 个实体在知识图谱中判断两者是否属于一种关系,并分别对属于或不属于 2 种情况给出概率。本文研究的重点是:基于对原模型改进后的 GraphDIVA 模型,在进行推理任务分析时,通过二分类模式由一个模型直接输出 2 种分类各自的概率,提升分类任务的准确性。GraphDIVA 模型的路径搜索过程对应了传播模式分析中舆情的核心传播过程路径的构建,而路径推理过程则对应了以核心传播路径为主干的传播关系子网的传播模式特征识别。

2.3 训练过程

原始的舆情传播知识图谱中只包含单向的传播关系,因此,在进行传播路径搜索与特征聚合时

也只能单向进行,无法进行传播溯源、邻居特征聚合等操作。因此,训练与测试数据集的搭建参考了文献[15]中建立 NELL995 数据的过程。

(from, relation, to)

(12)

(to, relation_{inv}, from)

(13)

建立数据集时,本文在原始知识图谱的基础上为每个三元组都创建了对应的反向关系三元组。

原始的 DIVA 模型与 GraphDIVA 模型只能对关系进行分类判断,因此,需要将传播模式对应为舆情事件中起点微博与 endpoint 微博之间的关系。本文定义的新闻类与娱乐类 2 种舆情传播模式分别建立了 political 与 entertainment 两类传播模式关系,并在始发舆情与转发舆情间建立对应的关系三元组作为训练与测试数据。

由于在搜集舆情时是通过关键词获取信息的,可以直接以关键词作为样本划分的依据。之后,对每类关键词中转发深度大于 1 的始发博文搜索位于最深转发位置的转发博文并保存传播路径。如果转发深度大于 5,则只搜索深度为 5 的博文,尽量减小深度差别带来的数据偏差。最终,本文构建了包含 18 633 个样本的数据集。样本的详细分布状况如表 9 所示。

表 9 样本分布状况

Table 9 Sample distribution status		
类型	新闻类/个	娱乐类/个
训练	4 534	9 441
测试	1 538	3 120

另一个与传统知识推理任务的差异,是舆情传播模式分析任务只关心从某个始发微博诞生的传播关系网属于哪种传播类型,而不关注于始发微博与某个中间转发微博之间属于哪种传播类型。因此,模型只在预训练阶段关注传播路径,保证路径搜索模块的基本搜索能力。之后的训练与测试过程不再对路径的终点进行检查,直接进入路径分类模块计算分类概率。实际上,路径搜索模块在本任务中的作用更类似核心传播路径的分析与生成,而路径推理模块则对应了图数据的结构分类功能。

3 实验结果及分析

本节针对舆情传播分类任务,完成了 GraphDIVA 模型和 DIVA 模型在不同训练样本中的传播模式分类正确率的对比实验。通过实验表明,本文方法针对小样本进行分类时,用较少的训练次数达到了较高的正确率。

3.1 实验参数

本文在生成传播图谱的嵌入向量时,使用了 Fast-TransX^[16]项目中的 TransE 实现。项目的代码由 C++ 编写,同时,从算法层面针对多线程运行进行优化,运行时只占用极少的内存。Fast-TransE 编码器编码参数如表 10 所示。表中:embedding_size 表示最终生成的嵌入向量宽度;nbatches 表示每轮迭代的批数量;threads 表示生成时使用的线程数;epochs 表示迭代次数;alpha 表示生成时的学习速率。

DIVA 模型与 GraphDIVA 模型在进行传播模式分析任务训练时的规模与参数如表 11 和表 12 所示。

表 10 Fast-TransE 参数

Table 10 Fast-TransE parameters

参数	数值
embedding_size	100
nbatches	1
threads	8
epochs	1 000
alpha	0.001

表 11 传播模式分析网络参数

Table 11 Propagation pattern analysis network parameters

参数	数值
finder_lstm_width	200
finder_mlp_width	200
reasoner_cnn_filter_size	64
reasoner_aggregate_width	200
reasoner_aggregate_neighbors	25
reasoner_lstm_width	200
reasoner_mlp_width	200

表 12 传播模式分析训练参数

Table 12 Propagation pattern analysis training parameters

参数	数值
guided_learn_rate	0.001
guided_max_epoch	25
guided_stop_growth	0.007 5
guided_max_path_width	5
unified_max_epoch	20
unified_posterior_learn_rate	0.01
unified_likelihood_learn_rate	0.000 25
unified_prior_learn_rate	0.001
unified_max_path_width	10

为了加快训练与测试速度,在训练时,先对训练样本进行随机排序,再选择前 100 个样本作为训练集;测试时,进行同样的随机排序操作,并选择前 500 个样本作为测试集。另外,为了突出

GraphDIVA 模型相比于原模型的性能提升,实验中额外测试了 2 个模型在前 25 个样本中进行训练后的性能表现。

3.2 实验结果

本文使用 DIVA 模型与 GraphDIVA 模型运行相同的数据并对比最终效果。100 个训练样本的实验结果如图 4 所示。结果显示,利用知识图谱与变分推理模型进行舆情传播模式分析是非常有效的一种技术方案,2 种模型使用 100 个训练样本在经过 4 轮训练后即可达到 90% 的分类正确率。GraphDIVA 模型相比于原始 DIVA 模型分类正确率从 93% 提升到 95%,并且在训练 10 轮后就达到了最高正确率;相比之下,DIVA 模型则需要 20 轮训练才能达到最佳效果。

25 个训练样本的测试结果如图 5 所示。结果显示,GraphDIVA 模型在经过 20 轮训练后就能达到 89.4% 的正确率,而 DIVA 模型只有 76%,实验表明 GraphDIVA 模型在减少训练次数、提升分类正确率方面具有更优的效果。

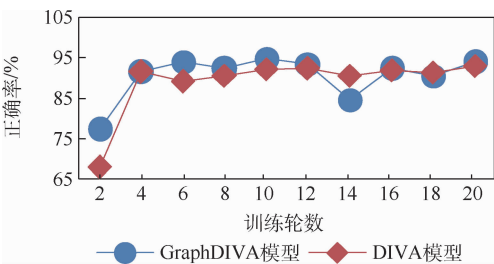


图 4 舆情传播模式分类 100 样本测试结果

Fig. 4 100 sample test results of public opinion propagation pattern classification

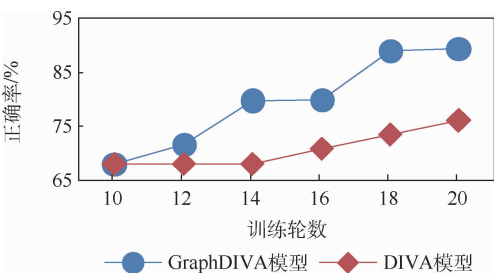


图 5 舆情传播模式分类 25 样本测试结果

Fig. 5 25 sample test results of public opinion propagation pattern classification

为了展示舆情分析模型的分析过程与效果,实验选取了“无锡高架桥坍塌”与“永不消逝的电波”2 个舆情事件作为新闻类与娱乐类传播模式的案例,分析后的结果分别对应图 6 与图 7,图中加粗的连接线表示模型识别出的核心传播路径。可以看出,虽然转发深度都是 9 层,模型从“无锡高架桥坍塌”的传播图谱中识别到的核心传播路

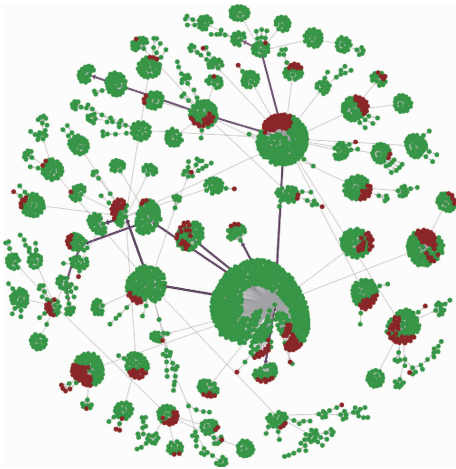


图 6 新闻类案例分析结果
Fig. 6 News case analysis results

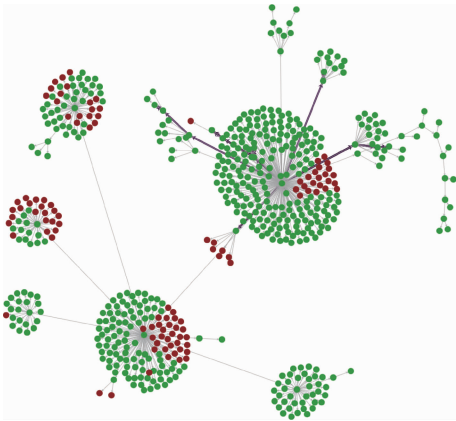


图 7 娱乐类案例分析结果
Fig. 7 Entertainment case analysis results

径多样性强于“永不消逝的电波”,且核心传播路径周边的传播子图结构特征有很大的差别,在路径末端仍能找出多关键点模式的结构,该实验体现了变分推理模型先构建核心传播路径,再对传播子网特征进行识别的分析流程的有效性。

4 结 论

- 1) 针对舆情传播分析任务场景,提出了舆情传播领域知识图谱结构定义,使用微博数据搭建了舆情传播知识图谱与舆情传播分析任务数据集。
- 2) 使用 GraphDIVA 模型进行舆情传播模式分类,围绕分析任务设计训练过程,并在自建数据集中完成了 GraphDIVA 模型和 DIVA 模型模式分类 100 样本和 25 样本测试的对比实验。GraphDIVA 模型能有效降低训练次数,提升分类的正确率,特别是在 25 样本测试中,模型在经过 20 轮训练后,分类正确率从 76% 提升到 89.4%。
- 综上所述,使用 GraphDIVA 模型进行传播模

式小样本分类时,在减少训练次数、提升分类正确率方面具有更优的效果。

参考文献 (References)

[1] 童亚拉. 突发群体性事件网络舆情信息传播复杂网络预测模型分析[J]. 微型电脑应用, 2011, 27(2) : 28-29.
TONG Y L. Analysis of forecasting module of information communication in mass emergency using theory of comple[J]. Microcomputer Applications, 2011, 27(2) : 28-29 (in Chinese).

[2] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016(4) : 589-606.
XU Z L, SHENG Y P, HE L R, et al. Review on knowledge graph techniques[J]. Journal of University of Electronic Science and Technology of China, 2016(4) : 589-606 (in Chinese).

[3] 王晰巍, 邢云菲, 赵丹, 等. 基于社会网络分析的移动环境下网络舆情信息传播研究——以新浪微博“雾霾”话题为例[J]. 图书情报工作, 2015, 59(7) : 14-22.
WANG X W, XING Y F, ZHAO D, et al. The study of network public opinion dissemination with social network analysis under the mobile environment: A case of “Haze” in Sina Micro-blog [J]. Library and Information Service, 2015, 59(7) : 14-22 (in Chinese).

[4] 崔树娟, 宾晟, 孙更新, 等. 基于大数据分析的多关系社交网络舆情传播模型研究[J]. 中南民族大学学报(自然科学版), 2018, 37(2) : 118-124.
CUI S J, BIN S, SUN G X, et al. Public opinion propagation model based on big data analytics in multiple relationships social network[J]. Journal of South-Central University for Nationalities(Natural Science Edition), 2018, 37(2) : 118-124 (in Chinese).

[5] 王兰成, 娄国哲. 大数据环境下涉军网络舆情的知识图谱服务研究[J]. 中华医学图书情报杂志, 2018, 27(4) : 4-9.
WANG L C, LOU G Z. Knowledge graph service for military network opinion in the big data era [J]. Chinese Journal of Medical Library and Information Science, 2018, 27(4) : 4-9 (in Chinese).

[6] 马哲坤, 涂艳. 基于知识图谱的网络舆情突发话题内容监测研究[J]. 情报科学, 2019, 37(2) : 33-39.
MA Z K, TU Y. Online emerging topic content monitoring based on knowledge graph[J]. Information Science, 2019, 37(2) : 33-39 (in Chinese).

[7] CHEN W, XIONG W, YAN X, et al. Variational knowledge graph reasoning[EB/OL]. (2018-10-23) [2020-09-01]. <https://arxiv.org/abs/1803.06581>.

[8] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. (2014-05-01) [2020-09-01]. <https://arxiv.org/abs/1312.6114v10>.

[9] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[C] // Advances in Neural Information Processing Systems, 2017: 1024-1034.

[10] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-22) [2020-09-01]. <https://arxiv.org/abs/1609.02907>.

[11] HOCHREITER S, SCHMIDHUBER J. Long short-term memory

[J]. *Neural Computation*,1997, 9(8):1735-1780.

[12] 娄国哲,王兰成. 基于知识图谱的网络舆情知识组织方法研究[J]. *情报理论与实践*,2019,42(1):58-64.
LOU G Z,WANG L C. Network public opinion knowledge organizing method based on knowledge map [J]. *Information Studies:Theory & Application*,2019, 42 (1) : 58-64 (in Chinese).

[13] 刘继,李磊. 基于微博用户转发行为的舆情信息传播模式分析[J]. *情报杂志*, 2013,32(7):78-81.
LIU J,LI L. Analysis of public opinion propagation mode based on repost behavior of microblog users [J]. *Journal of Intelligence*, 2013,32(7):78-81 (in Chinese).

[14] PRADIP K S,SHAIENDRA R,JONG H P. Multilevel learning based modeling for link prediction and users consumption preference in online social networks [J]. *Future Generation Computer Systems*,2019,93:952-961.

[15] XIONG W,HOANG T,WANG W Y. DeepPath:A reinforcement learning method for knowledge graph reasoning [EB/OL]. (2018-07-07) [2020-09-01]. <https://arxiv.org/abs/1707.06690v3>.

[16] LIN Y,LIU Z,SUN M,et al. Learning entity and relation embeddings for knowledge graph completion[C] // Twenty-ninth AAAI Conference on Artificial Intelligence,2015:2181-2182.

Classification of network public opinion propagation pattern based on variational reasoning

TANG Hongmei^{1,2}, TANG Wenzhong¹, LI Ruichen¹, WANG Yanyang^{3,4,*}, WANG Lihong⁵

(1. School of Computer Science and Engineering, Beihang University, Beijing 100083, China;

2. Science and Technology Projects Service Center of Xinjiang Uygur Autonomous Region, Urumqi 830000, China;

3. School of Aeronautic Science and Engineering, Beihang University, Beijing 100083, China;

4. Beihang Jiangxi Research Institute, Nanchang 330096, China;

5. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: With the rapid development of online social media, the analysis of the dissemination mode of public opinion information has become a research hotspot. Aiming at the problem of low classification accuracy of small sample data multi-path generation in the classification task of the network public opinion spreading pattern,the definition of the knowledge graph structure in the field of public opinion dissemination is proposed, builds a public opinion dissemination knowledge graph and public opinion dissemination analysis task data set based on Weibo data, uses the GraphDIVA model to classify public opinion propagation patterns, and conducts a 25-sample test experiment of public opinion propagation pattern classification in the self-built data set. The results show that, after 20 rounds of training, the classification accuracy rate of the model has increased from 76% to 89.4% . It can be seen that the GraphDIVA model has a better effect in reducing the number of training and improving the classification accuracy rate.

Keywords: public opinion propagation pattern; knowledge graph; knowledge graph reasoning; graph neural network; pattern analysis

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0422

基于启发式遗传算法的模糊测试 样本集优化方案

王志华*, 王浩帆, 程漫漫

(郑州大学 网络空间安全学院, 郑州 450002)

摘 要: 模糊测试作为当前最有效的漏洞挖掘方法, 不仅比其他漏洞挖掘技术更能应对复杂的程序, 而且可扩展性很强。在数据量相对较大的测试中, 模糊测试输入样本集存在质量低、冗余性高和可用性弱等问题。因此, 对模糊测试输入样本集进行研究, 提出了启发式遗传算法, 借助 0-1 矩阵, 通过启发式遗传算法对样本的执行路径进行选取和压缩, 从而获得优化后兼顾样本质量的样本集最小样本集合, 进而加快模糊测试的效率。实验结果表明: 在没有损失的情况下, 样本集精简后模糊测试的时间比精简前降低了 22%, 压缩率相比传统方案提升约 40%。

关 键 词: 漏洞挖掘; 模糊测试; 集合覆盖; 遗传算法; 样本集

中图分类号: TP399

文献标志码: A

文章编号: 1001-5965(2022)02-0217-08

随着信息与通信技术的快速发展, 信息技术改变了人们的生活方式。然而, 伴随而来的网络与信息安全问题也逐渐引起了社会各个领域的广泛关注。

软件漏洞挖掘技术^[1]在信息安全领域中占据了重要地位。模糊测试^[2]则是漏洞挖掘中较为有效的方法之一, 其是一种基于缺陷注入的自动化软件测试技术, 主要依赖于不断输入非预期的数据并对数据进行监控和观测的方式发现软件是否存在异常。传统的模糊测试技术并没有对初始样本集进行处理, 而是直接进行变异分析。样本的重复性会产生大量代码率低且无用的测试用例, 从而降低了模糊测试的效率。如何改进模糊测试的缺陷已经成为当前研究热点。考虑模糊测试样本集的特点, 可以对其进行优化, 从而精简样本集的数量, 提高样本集的质量, 进而提升模糊测试的效率。因此, 研究模糊测试样本集的高质

量性精简问题, 具有很大的理论意义和工程应用价值。本文对模糊测试样本集精简进行了研究, 借助 0-1 矩阵, 在遗传算法的基础上进行改进, 将贪心逼近算法与遗传算法结合, 提出了基于启发式遗传算法的样本集精简解决方案。实验结果表明, 本文方法是有效的。

1 相关工作

近年来, 众多学者从集合覆盖、机器学习、遗传算法等角度围绕模糊测试样本集的精简问题开展了大量的研究工作。马金鑫等^[3]设计了一种贪心逼近的求解算法来优化模糊测试的输入样本集, 在样本集覆盖率不变的前提下求解获取最小样本集合; Bhme 等^[4]把模糊测试问题建模为 Markov 模型, 并采用特定的策略引导 AFL 优先选择低频路径和变异频率低的文件作为文件进行变异; 随后 Bhme 等^[5]提出采用模拟退火算法对能

收稿日期: 2020-08-12; 录用日期: 2020-09-25; 网络出版时间: 2020-11-05 16:48
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201104.1906.007.html
基金项目: 河南省科技攻关项目(212102210408); 河南省高等学校重点科研项目计划(22A520041)
* 通信作者. E-mail: zhwang@zzu.edu.cn

引用格式: 王志华, 王浩帆, 程漫漫. 基于启发式遗传算法的模糊测试样本集优化方案[J]. 北京航空航天大学学报, 2022, 48(2): 217-224. WANG Z H, WANG H F, CHENG M M. Fuzzing testing sample set optimization scheme based on heuristic genetic algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 217-224 (in Chinese).

逼近特定目标位置的测试输入分配更高的能量,并优先选取高能量种子文件进行变异;Wang等^[6]提出了一种质量感知的测试用例优先级技术,优先输入高质量的种子,直接定义获取高质量的种子;唐泉^[7]利用污点分析和其他技术来获取数据的执行路径,对代码覆盖率相关字段进行基因编码,并执行遗传算法的选择变异过程,对危险操作相关字段执行边界值赋值,产生新的测试用例。

在实际处理过程中,采用集合覆盖技术解决样本集的精简或筛选时,精确算法能够找到集合覆盖问题的最优解,近似算法能够取得有质量保证的解,但其求解速度较慢,只能求解较小规模的实例。采用机器学习方法时,需要对数据进行大量的训练并分类,但由于模糊测试样本集的整体性特点,需要人工干预设定指标,反而降低了模糊测试的效率。使用传统的逼近算法可以实现对样本集的缩减,但是没有考虑样本在实际问题中的适用性。

2 基础理论和评价指标

2.1 基础理论

2.1.1 贪心逼近算法

贪心逼近算法^[3]是指利用有关算法解决一些实际问题的场景或精确度,且给出的解决方法是理论上的最优解。在样本集的精简中,贪心逼近算法指的是尽可能精确地给出样本的最小样本集,获取优质样本是贪心逼近算法的唯一标准。

2.1.2 遗传算法

遗传算法^[8]是一种基于生物进化原理构想出来的搜索最优解的仿生算法。其模拟基因重组与进化的自然过程,把待解决的问题的参数编程成二进制码或十进制码(也可编成其他进制码),即基因,若干基因组成一个染色体(个体)。对染色体进行类似于自然选择、配对交叉和变异运算,经过多次重复迭代(世代遗传)直至得到最后的优化结果。

遗传算法的步骤如下:

步骤1 初始化第0代种群 P_0 。

步骤2 对第*i*代种群 P_i 迭代执行步骤2.1~步骤2.5,直到满足停止准则:

2.1 计算 P_i 中每个个体的适应值,并按照适应值的大小对所有个体进行排序。

2.2 将 P_i 中适应值最佳的个体加入 P_{i+1} 。

2.3 按照适应值的大小排序在 P_i 中选择

2个父体。

2.4 按概率选择杂交算子或者变异算子对两父体进行遗传操作,将生成的个体加入 P_{i+1} 。

2.5 如果 P_{i+1} 的规模已经与 P_i 持平,则 $i \leftarrow i+1$ 并转步骤2,否则转步骤2.3。

步骤3 将最终种群中适应度最好的个体作为遗传算法的结果。

2.1.3 集合覆盖问题

集合覆盖问题(SCP)^[9]:在模糊测试中,任何样本集都可以转化为最小集合覆盖问题,而最小覆盖集为经典的NP-hard问题^[10],难以计算最优解。集合覆盖问题形式上可以描述如下:令 $A = (a_{ij})$ 为一个*m*行、*n*列的0-1矩阵。 $C = (c_j)$ 为一个*n*维整数向量。令 $M = \{1, 2, \dots, m\}$, $N = \{1, 2, \dots, n\}$ 表示矩阵 A 的行向量和列向量维数。值 $c_j (j \in N)$ 表示某一列的代价。不失一般性,假定 $c_j > 0, j \in N$ 。如果 $a_{ij} = 1$,则认为列 $j \in N$ 覆盖了行 $i \in M$ 。集合覆盖问题要求一个最小代价的子集合 $S \subseteq N$,这样每一行 $i \in M$ 至少被一列 $j \in S$ 覆盖。集合覆盖问题的一个自然的数学模型可以描述为: $v(SCP) = \min \sum_{j \in N} c_j x_j$, 服从于 $\sum_{j \in N} a_{ij} x_j \geq 1, i \in M, x_j \in (0, 1) (j \in N)$; 如果 $x_j = 1 (j \in S)$, 则 $x_j = 0$ 。

2.2 评价指标

1) 压缩率。对比样本精简前和样本精简后的测试样本数量,并计算压缩率来表示算法对样本集的优化程度。

2) 执行模糊测试所需要的时间。假设模糊测试初始样本集数量为*n*,模糊测试时间函数为*f(n)*,随着模糊测试数量*n*的增加,模糊测试执行时间的增长率与*f(n)*的增长率正相关。

3) 样本质量。主要由代码覆盖率为判断依据;代码覆盖率为测试的代码行数占总代码行数的比例。

3 基于启发式遗传算法的样本集优化

样本集的精简是将数量庞大的数据集进行缩减的过程。从本文所解决的问题出发,对于模糊测试集中的样本,样本的重复不仅表现在完全相同的样本,而且从变异的根源来说,样本间基本块的相互覆盖是导致变异产生大量重复测试用例的问题所在。因此,本文在进行样本选取时,规定优先选取包含代码基本块数最多的样本(代码块数多的样本在进行变异时可以得到更多测试用例),再按照样本性能依次选择,直到新样本集基

本代码块可以覆盖原始样本集的基本代码块为止。本文要定义的启发式算法是对染色体选择的算法,最终获取所求染色体集合。

3.1 染色体的表示

为了保证在样本集数量最少的基础上得到更优质的样本(基本代码块数多且变异能力好),在实现过程中,引入 0-1 矩阵^[11],向量 x 的元素是 0 或者 1,用一个 n 位的二进制串作为染色体结构, n 是矩阵 A 的列数,第 i 位上的值“1”意味着第 i 列被选择。

进行模糊测试时,每个程序都有自己所包含的基本代码块,根据基本块地址找到的内容就是对应的代码块。基本块地址信息的获取是比较方便的,故将基本块地址作为研究对象(每个基本地址块相当于遗传算法中的基因)。每个样本看作一个以基本代码块地址为元素的集合(样本相对于遗传算法中的染色体)。样本中如果存在某个基本地址块就用“1”表示,否则用“0”表示,所有的样本构成一个以 0 或 1 为元素的 0-1 矩阵。矩阵中的“1”即表示染色体的基因,每一列基因的集合相当于一个染色体。

3.2 启发式遗传算法改进染色体

选择的染色体都会有自己独特的基因存在,交叉重叠的基因也不计其数。在进行集合覆盖时,需要确定最小覆盖集合并保证该集合冗余最小。

在实际的遗传过程中,由于变异操作产生的染色体并不适合问题集合,本文不再对新个体的变异基因进行修复,而是将变异产生的新个体直接舍弃(由于原始数据较相似,出现新的基因的概率很小)。遗传算法执行后会保留所有产生的染色体,该启发式算法的作用包括以下 2 点:①消除因为基因重复产生的冗余;②选择更优质的染色体(包含的基因多且基因组合更丰富)。

启发式算法阐述如下:

1) 矩阵中的染色体和基因表示。

在 0-1 矩阵中,一列表示一个染色体,矩阵中的“1”表示某个染色体包含这个基因,“0”表示该染色体不包含该基因。算法使用性能代价比的优先思想:性能指的是染色体中某位基因所对应集合中的列包含“1”的个数,个数越多,则该列覆盖的行数就越多,说明该染色体的性能比较高。假设一组染色体使用遗传算法时产生的所有染色体集合为: $S_{all} = \{s_1, s_2, s_3, s_4, s_5\}$, $s_1 = \{2, 4, 6\}$, $s_2 = \{1, 6\}$, $s_3 = \{3, 5\}$, $s_4 = \{2, 3, 7\}$, $s_5 = \{1, 7\}$ (本文采用数字来表示其包含的基因及位置)。

2) 染色体的基因交换。

假设 A_1 和 A_2 是父代的 2 个染色体, B_1 和 B_2 是染色体交换之后所得的孩子染色体。可以覆盖矩阵 A 的集合,借助集合 C_1 和 C_2 , 用来存放基因没有覆盖的行号;集合 D_1 和 D_2 用来存放 B_1 和 B_2 包含的全部基因。例如, $A_1 = \{s_1, s_5\}$, $A_2 = \{s_2, s_3, s_4\}$ 集合, B_1 、 B_2 、 C_1 和 C_2 均为空集, $D_1 = \{1, 2, 3, 4, 5, 6, 7\}$, $D_2 = \{1, 2, 3, 4, 5, 6, 7\}$ 。

① 计算集合 A_1 和 A_2 中每个基因体的性能 η , 即每列所包含的“1”的个数,“1”越多,表示该染色体能覆盖行越多,其包含的基因就越多,其性能就越高。此时各个基因的性能为: $\eta_{s_1} = \eta_{s_4} = 3$, $\eta_{s_2} = \eta_{s_3} = \eta_{s_5} = 2$ 。

② A_1 和 A_2 中性能最高的染色体放入集合 B_1 , 统计该染色体包含的基因,并在集合 D_1 中将其删除,计算其未包含的基因,将这些基因的行号存入集合 C_1 ,再按照相同方法排列 A_1 和 A_2 剩下的基因到 B_1 ,最后将其他的基因放入 B_2 。

③ 在进行基因选择时,可能会出现一些基因性能一样的情况,因此在性能相同的基因中选取最优质的基因是算法设计的关键。例如,假设集合父辈 A_1 中有 2 个性能一样的基因: $s_1 = \{2, 4, 6\}$, $s_2 = \{1, 4, 6\}$, 当前集合 B_1 中包含一个基因 $s_3 = \{2, 3, 5, 6\}$, s_1 、 s_2 作为被选择对象, s_2 和 s_3 中相同的代码块要多于 s_1 和 s_3 的代码块,因此选择 s_2 比 s_1 更合适。除此之外,模糊测试样本在进行变异时,会根据自己的代码块随机变异生成测试用例。例如, $s_1 = \{2, 4, 6\}$, $s_2 = \{1, 4, 6\}$, 当前集合 B_1 中包含一个染色体 $s_3 = \{2, 3, 5, 6\}$ 。在染色体 s_1 中,2 号和 6 号代码块在 B_1 中已经出现,代码块进行变异后,相当于产生了一个新的基因组合。 s_3 中,当 3 号和 5 号基因不变异时,剩下的 2 号和 6 号的基因组合与染色体 s_1 在 4 号基因不产生变异的基础上是一样的效果。相比较,染色体 s_2 除了 4 号基因,1 号和 6 号基因在进行变异时会产生新的组合,进而产生新的测试用例。在实现消除冗余的基础上,在进行下一个染色体选择时,会出现性能相同的染色体,但其所包含的基因是不同的;在进行选择前,将等待选择的性能相同的染色体与集合 B_1 中已选的染色体做“差”,记录“差”集的元素个数,按照上述规则,元素个数多的“差”集与已选染色体不相同的个数就越多,那么对应的染色体即为下个候选染色体。

启发式遗传算法在原始种群的基础上借助于 0-1 矩阵获取所需染色体;在不需要改变遗传算法工作过程的前提下优化搜索条件,提升模糊测

试的效率。

3.3 启发式遗传算法实现

启发式遗传算法的实现过程在理论意义上与传统的遗传算法是没有区别的,同样需要重点考虑以下几个方面:①种群大小和种群初始化。种群样本集的大小可以人为控制,而不随机选择。②父代选择。在千万样本集中选择 2 个样本集作为父代,这是遗传算法仿生物的表现。③交叉率。确定父代在染色体的某个点进行交叉繁殖产生来后代。④变异率。产生后代的方法不仅依靠交叉,变异也是新物种最重要的来源。⑤精英比率。直接进入下一代而不参与基因交换的优秀个体的比例。⑥停止准则。传统的遗传算法需设定界限;样本集合覆盖问题需要设定其遗传的代数,或者某一个子代覆盖了所有测试路径时停止。

1) 种群大小和种群初始化。

使用遗传算法解决实际问题时,种群的大小对遗传算法所求得解的质量有很大影响。适当的种群数量可以提高算法的性能。种群数量过大,影响算法的效率;种群数量过小,整个算法过程起到的作用不大,反而将集合问题复杂化。在样本集精简问题中,本文算法没有对种群的大小进行太多的限制,数量根据实际情况自行定义,但是也会对不同种群数量进行不同的测试,以便粗略地获取较好的种群数量范围。

2) 父代的选择。

父代的选择是对种群中每个个体赋予的产生后代的机会。一般选择方法包括随机选择法、锦标赛选择法和轮盘赌注法。本文采用轮盘赌注法,其基本思想是:个体被选择的概率与其适应度大小成正比。具体实现操作如下:

① 计算出种群每个个体的适应度 $f_i (i = 1, 2, \dots, M, M$ 为种群大小)。

② 计算每个个体被遗传到下一代种群中的概率 $p_i = f_i / \sum_{i=1}^M f_i$ 。

③ 计算每个个体的概率分布 $q_i = \sum_{j=1}^i p(x_j)$ 。

④ 在 $[0, 1]$ 区间内产生一个均匀分布的伪随机数 r 。

⑤ 若 $r < q_1$, 则选择个体 1, 否则, 选择个体 k , 使得: $q_{k-1} < r \leq q_k$ 成立。

⑥ 重复④、⑤共 M 次。

3) 交叉率^[12]的选择。

交叉是产生新个体的主要方法,在被选中的

父代中确定一个交叉点,使双亲在交叉点进行基因的重组。交叉率是用来设定交叉池中参与交叉的染色体的个数,合理的交叉率可以保证交叉池中不断产生新个体,且不会产生过多的新个体,导致遗传秩序破坏。不同的交叉方法对应着不同的交叉率,当前主流的方案是采用自适应方法。

4) 变异率^[13]的选择。

变异率指的是一个种群中发生变异的基因数目与全体基因数目的比例。变异是产生新个体的另一种方式,可以通过设定随机变异的基因数也可以设定变异率来控制变异情况。在本文所研究的样本集精简中,变异可以将包含独特基因的染色体提前纳入到所设定的搜索规则中,以便于保证遗传算法的全局性。为此本文采用实验来设定合适的变异率,将最后一条含有独特基因的染色体混入其他染色体,并使用本文提出的启发式遗传算法来进行选择。

图 1 给出了变异率在 0.5、0.6 和 0.7 时待选基因随着初始染色体数量的增加被启发式遗传算法选入的时间变化。由图 1 可以得出,变异率为 0.6 时,花费时间最少。变异率过小会导致参与变异的染色体过少,并不能将含有独特基因的染色体提前录入集合;变异率过高将会使得参与变异的染色体过多,更容易产生一些非法的数据并且增大时间开销。因此,综合考虑后,本文启发式遗传算法选取的变异率为 0.6。

5) 精英比率。

当前种群中适应度最高的个体不参与交叉运算和变异运算,而是用来替换掉本代群体中经过交叉、变异等遗传操作后所产生的适应度最低的个体,本例中设置精英比率为 0.08。

6) 停止准则。

遗传算法在实际操作中都要经过多代进化,

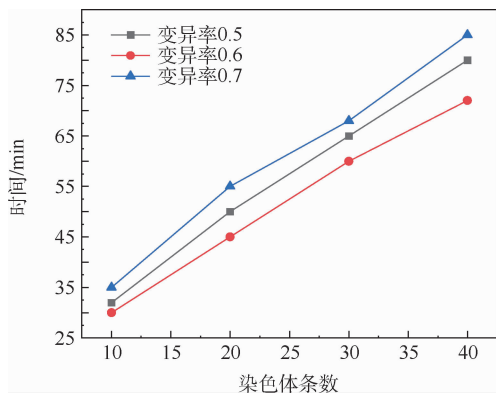


图 1 不同变异率选取染色体时间折线图

Fig. 1 Broken line graph of chromosome selection time with different mutation rates

直到适应值趋于稳定,或者找到最优解或者达到规定的遗传代数,进化过程结束。本文的样本集精简目标是在不降低代码覆盖的前提下尽可能降低测试用例的数量。当算法执行到特定迭代次数或选择出的测试用例不再发生变化时,算法停止。

4 实验设置与评估

4.1 实验准备

1) 实验环境。实验在 Windows 10 系统下,配置为 Intel Core i7-4790 CPU 3.6 GHz,8 GB 内存,使用 peach 等工具采用启发式遗传算法对模糊测试样本集进行精简优化。

2) 测试环境与目标程序。在模糊测试实验中,本文选取 peach^[14]作为模糊测试工具:①安装方便;②有开放的使用说明;③不同的命令可实现多种操作,在获取样本的执行路径时无需再借助于其他工具。同时选取 mspaint、pdfium、VLC Player 作为目标程序。

3) 数据来源。本文实验所需的 jpg、PDF、AVI 类型数据来源于互联网开源数据集。此数据集中包括很多重复数据,并且不同数据之间存在交叉;其将增加模糊测试样本集录入的时间,并使 peach 在变异时产生更多重复甚至没有任何利用价值的测试用例。

4.2 实验方法与分析

3.3 节阐述了本文方案的算法评价指标,通过时间复杂函数来展现算法随着初始数据的变化而变化的程度,而这些表现可以多种角度来展示。为了更直观地展现算法的有效性,设计了以下实验,并以 jpg 类型数据实验过程进行分析。

1) 样本集数量。

启发式遗传算法在 0-1 矩阵的基础上操作,将矩阵中的元素“1”看作是遗传算法中的染色体。在遗传算法过程中,通过父代染色体的变异和交叉产生新的染色体,父代的选择也是样本精简的一个过程,只选择性能好的染色体作为父代,直至进化到停止准则要求时间。实验中,对初始样本集做了定向的选择。jpg1:随机选择;jpg2:3 组 jpg1;jpg3:随机选择,jpg4:随机选择。精简样本集的数量与初始样本集的数量关系在表 1 中给出。由前 2 组数据可以看出,第 2 组数据在启发式遗传算法后数量明显减少,jpg2 包含 3 组 jpg1,jpg2 的精简样本集数量与第 1 组一样,启发式遗传算法在某种程上可以对样本进行精简,但是 jpg2 是特殊的数据组,不能体现算法的实用性。jpg3 随机选取的 2 000 个样本集和 jpg4 随机选择

表 1 样本集数量

Table 1 Number of sample sets

样本集合名	初始样本集数量	精简样本集数量	压缩率/%
jpg1	1 000	409	59
jpg2	3 000	409	86
jpg3	2 000	706	65
jpg4	4 000	1 332	67

的 4 000 个样本集,其经过启发式遗传算法优化后获得相应精简样本集,即启发式遗传算法并不会受选取样本的影响。从 jpg2 和 jpg3 两组数据可以看出,该启发式遗传算法可以减少样本集的数量;从 jpg1、jpg3 和 jpg4 的初始样本集数量和精简样本集的数量进行观察,假设存在精简比例,那么 jpg3 的精简样本集数量应该是 818,jpg4 的精简样本集数量应该是 1 636,而实际精简样本集数量却比期望低。这是因为:选取样本集数越大,该样本集所包含“基因”越多,其可覆盖“染色体”就越多,剩余不能覆盖的“染色体”就越少,那么精简样本集数量就越少。

2) 模糊测试时间。

为了验证本文启发式遗传算法的有效性及能否提升模糊测试效率,对整体模糊测试时间进行统计。将表 1 中所采用的 4 组数据进行启发式遗传算法优化后进行模糊测试,其花费的时间如表 2 所示,通过模糊测试时间的变化来说明该启发式遗传算法是否可以提升模糊测试效率。样本数量与测试用例个数的对比情况如图 2 所示。在一定程度上,测试用例的产生主要取决于原始样本的最小样本集合。模糊测试时间与样本数量情况如图 3 所示。精简后的样本模糊测试时间低于未精简样本的模糊测试时间;样本集精简后模糊测试的时间比精简前降低了 22%。本质上,模糊测试用例的数量由输入样本集的数量来决定,而模糊测试用例通过启发式遗传算法去除冗余可以提升模糊测试的效率。

3) 样本质量。

每个样本都有对应的代码块,根据不同的代码块变异而产生测试用例。在未精简时,重复样本产生无用测试用例。在时间一定的前提下,输入同一组初始种群,对产生的测试用例进行标记以保证获取的测试用例的不重复性,统计精简前后变异产生的测试用例的数量。本文从 jpg3 和 jpg4 中各取一个子集,名为 Test1 和 Test2。实验测试用例数据情况如表 3 所示。测试时间一定时,未精简的样本集短时间内产生的测试用例大于精简后的样本集所产生的测试用例,但其标记

后的测试用例个数小于精简后的标记个数。样本集在经过启发式遗传算法后,其样本的质量有所提升;在相同的时间内,其产生的测试用例存在的重复性小,同时精简前后的代码覆盖率没有发生变化,算法兼顾了样本质量。

4) 对比实验。

通过比较文献[3,15]提出的算法,观察不同模糊测试样本集优化方案的处理效果。文献[3,15]对比实验按照文献原文中设定参数进行精简优化,对千兆级样本数据进行处理,实验数据如

表4所示。通过表4可以观察出,使用文献[3,15]和本文所述样本集精简方案对初始样本集进行优化后,样本集中数量有一定下降,且代码覆盖率都维持在初始样本水平,保证了样本质量。但进一步观察,结合图4所示,贪心逼近算法运算速度较快,但需多次迭代方能达到和其他算法相同的精简数量,因此其速度却远远要慢于启发式遗传算法的速度;并且如表4所示,本文所述方案处理过的样本集中数量少于文献[3,15]提出的方案,说明经过本文所述方案优化后的样本集更加有效。

表 2 模糊测试时间

Table 2 Fuzzy testing time

初始样本集数	初始样本集 测试用例个数	初始样本集模糊 测试时间/min	精简样本 集数	精简样本集 测试用例个数	精简样本集模糊 测试时间/min
1 000	150 538	2 145	409	61 574	1 984
3 000	469 614	6 568	409	64 217	5 656
2 000	323 846	4 371	706	114 312	3 944
4 000	626 024	8 423	1 332	206 645	6 443

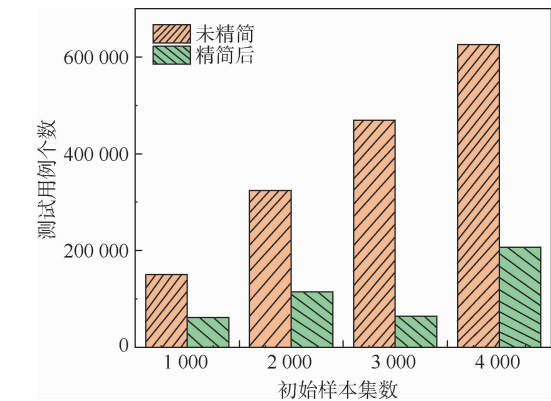


图 2 测试用例个数与样本数量

Fig. 2 Number of test cases and number of samples

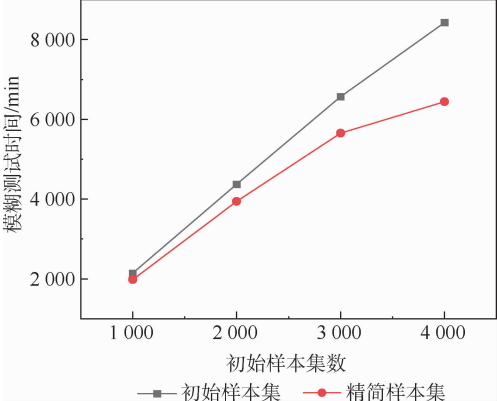


图 3 模糊测试时间与样本数量

Fig. 3 Fuzzy testing time and sample size

表 3 测试用例数据

Table 3 Test case data

样本 集合名	未精简样本集 产生的测试 用例个数	未精简样本集 产生的标记 测试用例个数	未精简代 码覆盖率/%	精简样本集 产生的测 试用例个数	精简样本集 产生的标记 测试用例个数	精简代码 覆盖率/%
Test1	3 075	2 356	37	2 954	2 879	37
Test2	5 974	4 517	32	5 804	5 691	32

表 4 多种方案测试样本数量和覆盖率

Table 4 Test sample size and coverage rate for multiple schemes

软件名称	样本数量				代码覆盖率,压缩率/%			
	初始样本	文献[3]方案	文献[15]方案	本文优化方案	初始样本	文献[3]方案	文献[15]方案	本文优化方案
mspaint	10 000	4 198	6 158	3 684	37,0	37,58	37,38	37,63
pdfium	10 000	4 517	7 141	5 804	25,0	25,54	25,28	25,41
VLC Player	10 000	3 759	4 106	3 406	29,0	29,62	29,58	29,65

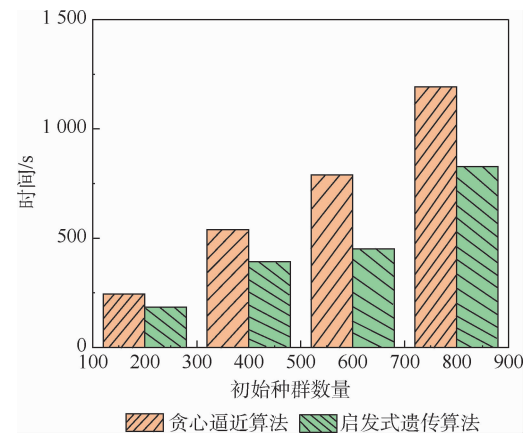


图 4 不同算法产生相同精简样本集所需时间对比

Fig. 4 Comparison of time required for different algorithms to produce the same simplified sample set

5 结 论

1) 启发式遗传算法在遗传算法上进行改进并且应用到集合覆盖问题中,通过染色体来表示集合中的覆盖情况,以高质量的染色体为背景,对染色体进行启发式改进,使之在种群中随时进化,从而精简优化样本集。该方案比传统方案压缩率提升约 40%,展示了该算法的有效性。

2) 本文在进行优质染色体选择过程中使用 0-1 矩阵与集合覆盖的方式来完成优质染色体的选取工作,并且改进了基因交叉的规则,加快模糊测试速度,同时能够兼顾样本质量和处理时间,测试时间降低了约 22%,比传统方式更加高效。

参考文献 (References)

[1] 邹权臣,张涛,吴润浦,等. 从自动化到智能化:软件漏洞挖掘技术进展[J]. 清华大学学报(自然科学版),2018,58(12):1079-1094.

ZOU Q C,ZHANG T,WU R P,et al. From automation to intelligence: Progress in software vulnerability mining technology [J]. Journal of Tsinghua University (Natural Science Edition),2018,58(12):1079-1094 (in Chinese).

[2] 张雄,李舟军. 模糊测试技术研究综述[J]. 计算机科学,2016,43(5):1-8.

ZHANG X,LI Z J. Research review of fuzzy testing technology [J]. Computer Science,2016,43(5):1-8 (in Chinese).

[3] 马金鑫,张涛,李舟军,等. Fuzzing 过程中的若干优化方法[J]. 清华大学学报(自然科学版),2016,56(5):478-483.

MA J X,ZHANG T,LI Z J,et al. Optimization methods in the Fuzzing process [J]. Journal of Tsinghua University (Natural Science Edition),2016,56(5):478-483 (in Chinese).

[4] BHME M,PHAM V T,ROYCHOUDHURY A. Coverage-based greybox fuzzing as Markov chain [C]//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York:ACM,2016:1032-1043.

[5] BHME M,PHAM V T,NGUYEN M D,et al. Directed greybox fuzzing [C] // Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York:ACM,2017:2329-2344.

[6] WANG S,NAM J,TAN L. QTEP:Quality-aware test case prioritization [C] // Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. New York:ACM,2017:523-534.

[7] 唐泉. 基于动态污点分析的反馈式模糊测试改进方法[J]. 信息安全研究,2019,5(2):145-151.

TANG X. Improved method of feedback fuzzy test based on dynamic stain analysis [J]. Information Security Research,2019,5(2):145-151 (in Chinese).

[8] 冯智莉,易国洪,李普山,等. 并行化遗传算法研究综述[J]. 计算机应用与软件,2018,35(11):1-7.

FENG Z L,YI G H,LI P S,et al. Research review of parallel genetic algorithm [J]. Computer Applications and Software,2018,35(11):1-7 (in Chinese).

[9] 林耿,关键. 自适应 memetic 算法求解集合覆盖问题[J]. 浙江大学学报(理学版),2016,43(2):168-174.

LIN G,GUAN J. Adaptive memetic algorithm for solving set coverage problem [J]. Journal of Zhejiang University (Science Edition),2016,43(2):168-174 (in Chinese).

[10] ALYAHYA K,ROWE J E. Landscape analysis of a class of NP-hard binary packing problems [J]. Evolutionary Computation,2019,27(1):47-73.

[11] 赵颖闻,王峻,郭茂祖,等. 基于 0-1 矩阵分解的蛋白质功能预测[J]. 中国科学:信息科学,2019,49(9):1159-1174.

ZHAO Y W,WANG J,GUO M Z,et al. Prediction of protein function based on 0-1 matrix decomposition [J]. Scientia Sinica:Information Science,2019,49(9):1159-1174 (in Chinese).

[12] 张瑜,娄卉芳,文良浩,等. 一种改进的遗传算法交叉策略[J]. 湖南科技大学学报(自然科学版),2012,27(1):94-97.

ZHANG Y,LOU H F,WEN L H,et al. An improved genetic algorithm crossover strategy [J]. Journal of Hunan University of Science and Technology (Natural Science Edition),2012,27(1):94-97 (in Chinese).

[13] 王春阳,赵玉庆,谢金兴,等. 遗传算法变异算子的改进[J]. 山东农业大学学报(自然科学版),2019,50(5):898-901.

WANG C Y,ZHAO Y Q,XIE J X,et al. Improvement of genetic algorithm mutation operator [J]. Journal of Shandong Agricultural University (Natural Science Edition),2019,50(5):898-901 (in Chinese).

[14] 伊胜伟,张骞斌,谢丰,等. 基于 Peach 的工业控制网络协议安全分析[J]. 清华大学学报(自然科学版),2017,57(1):50-54.

YI S W,ZHANG C B,XIE F,et al. Security analysis of industrial control network protocol based on Peach [J]. Journal of Tsinghua University (Natural Science Edition),2017,57(1):50-54 (in Chinese).

[15] 李谦,刘嘉勇. 基于 Fuzzing 技术的样本集精简优化研究[J]. 网络安全技术与应用,2017(1):62-64.

LI Q,LIU J Y. Research on simplification and optimization of

Fuzzing testing sample set optimization scheme based on heuristic genetic algorithm

WANG Zhihua* , WANG Haofan , CHENG Manman

(School of Cyberspace Security , Zhengzhou University , Zhengzhou 450002 , China)

Abstract: As the most effective method of vulnerability mining at present, fuzzy testing not only is more capable of dealing with complex programs than other vulnerability mining techniques, but also has strong scalability. In the fuzzy testing with a large number of data, the input sample set has the problems of low quality, high redundancy and weak availability. Therefore, we study the input sample set of fuzzy testing, and propose a heuristic genetic algorithm. With the help of the 0-1 matrix, the execution path of the sample is selected and compressed through the heuristic genetic algorithm, so as to obtain the smallest sample set that takes into account the sample quality after optimization, thereby speeding up the efficiency of fuzzy testing. The experimental results show that, without loss, the fuzzy testing time after the sample set is simplified is reduced by 22% compared with that before the sample set is simplified, and the compression rate is increased by about 40% compared with the traditional scheme.

Keywords: vulnerability mining; fuzzy testing; set coverage; genetic algorithm; sample set

Received: 2020-08-12; **Accepted:** 2020-09-25; **Published online:** 2020-11-05 16:48
URL: kns.cnki.net/kcms/detail/11.2625.V.20201104.1906.007.html
Foundation items: Henan Provincial Department of Science and Technology Project (212102210408); Henan Provincial Key Scientific Research Project (22A520041)
*** Corresponding author.** E-mail: zhwang@zzu.edu.cn

http://bxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0567

Android 恶意软件检测低冗余特征选择方法

郝靖伟¹, 潘丽敏¹, 李蕊¹, 杨鹏^{2,*}, 罗森林¹

(1. 北京理工大学 信息与电子学院, 北京 100081; 2. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 针对 Android 恶意软件检测特征选择中,对类间具有相同频率分布的特征过度关注而导致特征冗余问题,提出了一种 Android 恶意软件检测低冗余特征选择方法。利用 Mann-Whitney 检验方法选择出存在频率分布偏差的特征;通过外观比率间隔算法量化偏差程度和特征出现频率剔除低偏差和整体软件中低频使用的特征;结合粒子群优化算法和分类器检测效果得到最优特征子集。使用公开数据集 DREBIN 和 AMD 进行实验,实验结果显示,在 AMD 数据集上选择出了 294 维特征,进行特征选择后 6 种分类器的检测准确率提高了 1% ~ 5%,在 DREBIN 数据集上选择出了 295 维特征,少于 4 种对比方法,且进行特征选择后 6 种分类器的检测准确率提高了 1.7% ~ 5%。实验结果表明,所提方法能够降低 Android 恶意软件检测中特征的冗余性,提升恶意软件的检测准确率。

关 键 词: Android 恶意软件检测; 特征选择; Mann-Whitney 检验; 粒子群优化算法; 外观比率间隔算法

中图分类号: V219; TP317

文献标志码: A **文章编号:** 1001-5965(2022)02-0225-08

中国互联网络信息中心在 2019 年发布的《第 44 次中国互联网络发展现状统计报告》^[1]显示,中国移动互联网用户数已达 7.88 亿,手机互联网用户比例高达 98.3%。国际数据中心发布的报告^[2]显示,2019 年 Android 系统的市场份额为 87%,占据了智能手机操作系统市场的主导地位。该系统具备的灵活性和开放性给开发者和用户带来了极大的便利,如开发者能够自由开发并上传应用程序到应用市场,用户可以从应用市场中随意下载应用程序,但同时也使 Android 系统和用户容易遭受恶意软件的攻击,因此,研究 Android 恶意软件检测方法是移动端操作系统安全防护领域的重要内容之一。

机器学习涉及概率论、统计学等多门学科,是近年来兴起的最热门的多领域交叉学科之一,现已被广泛应用到 Android 恶意软件检测领域。基

于机器学习的 Android 恶意软件检测方法不需要过多人工干预,并且可随着恶意软件数据量的增加实现方法的自我改进,具有自动化和高精度的特点。特征选择是基于机器学习的 Android 恶意软件检测方法的重要组成部分。通过特征选择能够从高维特征空间中挖掘出用于检测恶意软件的关键特征。

针对 Android 恶意软件检测的特征选择方法,主要分为基于过滤评价策略的特征选择方法和基于包装评价策略的特征选择方法。

1) 基于过滤评价策略的恶意软件检测特征选择方法。主要包含信息增益、卡方检验、频率统计等方法。其中,卡方检验、信息增益等方法计算特征与检测结果的相关度,再对特征进行排序,选择相关度高的特征。例如,Yerima 等^[3]提出利用信息增益对恶意软件原始权限特征集进行特征选

收稿日期: 2020-09-30; 录用日期: 2020-12-18; 网络出版时间: 2021-03-02 14:23
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210302.0900.001.html
基金项目: 国家 242 信息安全计划(2019A012); 工信部 2020 年信息安全软件项目(CEIEC-2020-ZM02-0134)
* 通信作者. E-mail: yp@cert.org.cn

引用格式: 郝靖伟, 潘丽敏, 李蕊, 等. Android 恶意软件检测低冗余特征选择方法[J]. 北京航空航天大学学报, 2022, 48(2): 225-232. HAO J W, PAN L M, LI R, et al. Low redundancy feature selection method for Android malware detection[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 225-232 (in Chinese).

择,选择出信息增益最大的 20 个特征作为特征集训练朴素贝叶斯恶意软件检测器,检测准确率达 92%。Pehlivan 等^[4]提出通过卡方检验算法从 97 个权限特征中选择出 25 个特征,利用该特征集训练随机森林(RF)模型,检测准确率达到 94.9%。Wang 等^[5]先通过互信息、相关系数和 T-test 检验方法分别对权限特征进行排序,再使用顺序正向选择算法和主成分分析(PCA)算法识别有风险的权限子集,使用支持向量机(SVM)和随机森林等评估选择出的权限特征的检测效果,实验结果显示,该方法能够达到 94.62% 的检测率及 0.6% 的误报率。频率统计的方法通过计算特征在软件中出现的频率,选择频率高的特征。但是,Cen 等^[6]在利用该方法进行 Android 软件特征选择时研究发现,有些低频率出现的特征(如 SMS 相关的权限)对检测效果贡献较高不被选择,有些高频率出现的特征(如 toString()在恶意软件和良性软件中同时出现)对检测效果贡献度较低被选择,即该特征选择方法忽略了特征在类间的频率分布,容易选择出冗余特征。

2) 基于包装评价策略的恶意软件检测特征选择方法。Pehlivan 等^[4]提出一种基于权限特征的恶意软件检测方法,利用随机森林进行特征选择,选择出 25 维特征,达到了 89.32% 的准确率。Fest^[7]利用 Freqensel 算法对权限、API 等特征进行特征选择,得到 92% 的准确率和召回率。2018 年,Tao 等^[8]提出一种基于与权限相关的 API 特征的恶意软件检测方法,使用随机森林进行特征选择,选择出 50 个高度敏感的 API,得到了

98.24% 的 F_1 值。Li 等^[9]经过多层特征选择,从 135 维权限特征中选择出 22 维特征,利用该特征集构建基于支持向量机的检测模型,实现对 Android 恶意软件的检测。特征选择能够从高维特征空间中挖掘出 Android 恶意软件检测的关键特征,提高检测准确率。现有的 Android 恶意软件特征选择方法忽略特征在类间的频率分布,有些高频率出现的特征(如用户界面设计“android.view.ViewGroup.addView”在恶意软件和良性软件中同时出现)对检测效果贡献度较低,却被过度估计重要性,成为检测 Android 恶意软件的特征,导致特征子集冗余性较高,影响检测效果。

针对现有特征选择方法中对类间相同频率分布的特征过度关注导致特征冗余性较强的问题,本文提出一种 Android 恶意软件检测低冗余特征选择方法。首先,分离类间存在频率分布偏差的特征,然后,量化偏差程度和频率以剔除低偏差和低频特征,并结合分类器实际分类效果选择出特征子集,降低恶意软件检测特征集的冗余性。

1 算法原理

1.1 原理框架

Android 恶意软件检测低冗余特征选择方法的核心思想是:通过分离类间存在频率分布偏差的特征,量化偏差程度和特征出现频率以剔除低偏差和低频特征,并结合分类器实际分类效果选择出特征子集。其基本原理框架如图 1 所示。

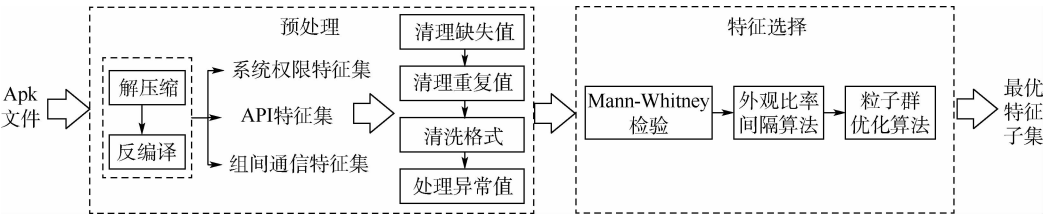


图 1 Android 恶意软件检测低冗余特征选择方法原理框架

Fig.1 Framework of low redundancy feature selection method for Android malware detection

该方法包括 2 个主要的功能模块:预处理模块和特征选择模块。预处理模块对输入的 Apk 文件进行解压缩和反编译得到 Smali 源代码和 Android Manifest.xml 文件,再从中提取出 3 种类型的特征集,并对提取出的特征集进行缺失值清理、重复值清理、格式清洗和异常值处理等处理;特征选择模块对处理后的特征集进行 Mann-Whitney 检验,提取出存在频率分布偏差的特征集,再利用外观比率间隔算法对特征的偏差程度和出现

频率进行量化并选择,最终通过粒子群优化算法结合分类器实际检测效果得到最优特征子集。

1.2 预处理

首先,使用 Androguard 工具^[10]对 Android 应用程序的 Apk 文件进行解压缩。其次,反编译解压缩后的文件,得到 Smali 源代码及 AndroidManifest.xml 文件。然后,解析 Smali 源代码中的 invoke-virtual、invoke-direct、invoke-static 和 invoke-super 四个指令的汇编代码,得到 API 特征 $S_{Api} =$

$\{p_1, p_2, \cdots, p_m\}$ 。其中, m 为应用程序包的数量。当 p_m 出现在 Apk Smali 源代码中时, 其值为 1, 否则为 0。最后, 解析 AndroidManifest.xml 文件, 从文件中获取访问控制机制的特征和组间通信的特征, 如权限、系统组件包含的 Intent 元素等。分别从 AndroidManifest.xml 文件和 Smali 源代码中提取出 24 个权限特征、45 个系统 Intent 特征及 620 个敏感 API 作为原始特征集。初始特征的具体信息如表 1 所示。

得到原始特征集后, 对数据进行清理缺失值、清理重复值、清洗格式及处理异常值的操作。清理缺失值主要是将数据集中有属性值缺失的数据删除, 如若某个权限的数据值不存在, 则直接删除该应用程序的所有数据; 清理重复值主要是以应用程序的 ID 作为标准, 删除所有具有相同 ID 的数据, 使得数据集中每个 ID 仅有一条数据; 清洗格式主要是删除数据集中的不合法字符, 如“-”等, 修改不正确的数据格式, 或将权限数量值由“科学计数”格式修改为“文本”格式等; 处理异常值主要是识别数据集中的离群点及异常点并删除该条数据, 如某个 ID 中权限特征值为 10 000 的数据。

经过数据预处理后, 得到一个无空缺属性值、无重复应用程序、无异常属性值及数据格式正确统一的 Android 恶意软件特征集, 以便后续步骤使用。

表 1 初始特征信息
Table 1 Initial feature information

特征类型	数目	特征
系统权限	24	READ CALENDAR
		WRITE CALENDAR
		CAMERA
		READ CONTACTS
		WRITE CONTACTS
		
API	620	android/net/ConnectivityManager; startUsingNetworkFeature
		android/net/wifi/WifiManager; enableNetwork
		android/net/wifi/WifiManager; disconnect
		android/net/wifi/WifiManager; setWifiEnabled
		
组间通信 Intent	45	Android.intent.action.MAIN
		Android.intent.action.VIEW
		Android.intent.action.ATTACH_DATA
		Android.intent.action.EDIT
		Android.intent.action.PICK
		

1.3 特征选择

1.3.1 Mann-Whitney 检验

经过预处理后得到包含系统权限、API 和组间通信 Intent 特征的原始特征集。接下来使用 Mann-Whitney 检验^[11]来分析这些原始特征集中的特征在恶意和良性软件中的频率分布是否有显著不同。

Mann-Whitney 检验方法是一种非参数秩和假设检验, 对独立样本进行的一种不要求正态分布的 T-test 检验方法, 能够对来自除了总体均值以外完全相同的总体, 检验是否具有显著差异。首先, 假设 2 个独立样本之间没有差异, 成立则 H_0 , 不成立则 H_1 。该方法的具体步骤如下:

步骤 1 检验的 2 组独立样本 n_1 、 n_2 先进行混合, 并根据数据大小升序排列并编排等级 (秩 rank), 遇到相同的数据时, 等级值相等, 为编排等级前的平均值。

步骤 2 分别求出 2 个样本的等级和 R_1 、 R_2 。

步骤 3 Mann-Whitney 检验统计量 U_1 、 U_2 的计算公式如下:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}$$

(1)

$$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

(2)

式中: U_1 、 U_2 中的最小值用于与显著检验阈值 U_α (查 Mann-Whitney Table 可得具体值) 相比较, 如果 $U_{\min} < U_\alpha$ 时则拒绝 H_0 , 接受 H_1 , 表明两样本之间存在差异。

分别假设在恶意软件和良性软件中第 i 个特征集没有显著不同 (不存在分布偏差), 拒绝零假设的阈值 α 为 0.05。再将恶意和良性的软件分为 2 组, 输入矩阵如表 2 所示。如果一个应用软件 (恶意或者良性) 包含某个特征, 则该样本的值 B_i 或 M_i 为 1, 反之为 0。在一个检验中, 2 组样本集代表 1 个特定权限是否被恶意或者良性软件使用。Mann-Whitney 检验的输出为拒绝假设或接受假设。当 Mann-Whitney 检验的 p 值大于阈值时则拒绝假设, 当 Mann-Whitney 检验的 p 值小于阈值时则接受假设。

表 2 Mann-Whitney 检验输入矩阵
Table 2 Mann-Whitney test input matrix

数据集	特征 f_i	数据集	特征 f_i
恶意软件 1	0	良性软件 1	0
恶意软件 2	0	良性软件 2	0
恶意软件 3	0	良性软件 3	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
恶意软件 n	B_i	良性软件 m	M_i

1.3.2 外观比率间隔算法

通过 Mann-Whitney 检验得到存在分布偏差的特征集 F ,接下来利用外观比率间隔算法量化特征在恶意软件和良性软件中的分布偏差程度,以及特征在良性软件和恶意软件中的出现频率,本文目标是:选择出在整体软件中高频出现且在恶意软件和良性软件中分布偏差程度大的特征。假设比例越小,一个特征在良性应用和恶意软件之间的外观比例差距越大。因此,显示小比例的特征对于检测恶意软件是有用的。外观比率间隔算法主要步骤如下:

步骤 1 分别计算特征集 F 中的特征 f 在良性软件和恶意软件中的外观比率 $E(f)$ 。特征在良性软件的外观比率 $E(f)_{\text{benign}}$ 和恶意软件的外观比率 $E(f)_{\text{malicious}}$ 计算公式如下:

$$E(f)_{\text{benign}} = -\log_2 \frac{n(B_f^0)}{n(B_f)} \tag{3}$$

$$E(f)_{\text{malicious}} = -\log_2 \frac{n(M_f^0)}{n(M_f)} \tag{4}$$

式中: $n(B_f^0)$ 和 $n(M_f^0)$ 分别为特征在良性软件和恶意软件中 0 值的数量; $n(B_f)$ 和 $n(M_f)$ 分别为良性软件和恶意软件的数量。

步骤 2 计算特征 f 的外观比率比例 $E(f)_{\text{proportion}}$ 和外观比率比例差异 $E(f)_{\text{difference}}$,公式如下:

$$E(f)_{\text{proportion}} = \frac{\min(E(f)_{\text{benign}}, E(f)_{\text{malicious}})}{\max(E(f)_{\text{benign}}, E(f)_{\text{malicious}})} \tag{5}$$

$$E(f)_{\text{difference}} = |E(f)_{\text{benign}} - E(f)_{\text{malicious}}| \tag{6}$$

步骤 3 特征选择。若特征 f 的外观比率比例 $E(f)_{\text{proportion}}$ 不为 0,则判断外观比率比例 $E(f)_{\text{proportion}}$ 是否小于阈值 $T_{\text{proportion}}$,若小于阈值,则该特征为只在良性或恶意软件中频繁出现的特征,该特征被选择进入特征子集,若大于阈值,则该特征为只在良性或恶意软件中较少出现的特征,不被选择进入特征子集。若外观比率比例 $E(f)_{\text{proportion}}$ 为 0,则判断外观比率比例差异 $E(f)_{\text{difference}}$ 是否小于阈值 $T_{\text{difference}}$ 。若小于阈值,则该特征为只在良性或恶意软件中频繁出现的特征,该特征被选择进入特征子集 F' ,若大于阈值,则该特征为只在良性或恶意软件中较少出现的特征,不被选择进入特征子集 F' 。外观比率间隔算法伪代码如下。

算法 1 外观比率间隔算法。

输入: $F, n(B_f^0), n(M_f^0), n(B_f), n(M_f)$ 。

输出: F' 。

For i to $F.size()$ do

 计算 $E(f)_{\text{benign}}, E(f)_{\text{malicious}}, E(f)_{\text{proportion}}, E(f)_{\text{difference}}$

 If $E(f)_{\text{proportion}}$ 不为 0 then

 If $E(f)_{\text{proportion}}$ 小于阈值 $T_{\text{proportion}}$ then

$F \leftarrow f_i$

 Else: $F \leftarrow f_i$

 Else:

 If $E(f)_{\text{difference}}$ 是否小于阈值 $T_{\text{difference}}$ then

$F \leftarrow f_i$

 Else: $F \leftarrow f_i$

1.3.3 粒子群优化算法

通过外观比率间隔算法提出低频和低偏差的特征后,接下来利用粒子群优化算法选择出用于 Android 恶意软件检测的特征子集。

首先,将粒子群随机初始化即随机选择特征子集。粒子的位置即特征子集 $X_i = \{x_1, x_2, \dots, x_N\}, x_i \in \{0, 1\}, i = 1, 2, \dots, N$ 。每个粒子的位置代表一个特征子集。其中, N 为特征的总数, $x_i = 1$ 代表 i 位置的特征被选择, $x_i = 0$ 代表对应位置的特征不被选择。初始化时,每个粒子位置在 $[0, 1]$ 之间。

然后,计算适应度 (Fitness),公式如下:

$$\text{Fitness} = \alpha \gamma_R(D) + \beta \frac{|C| - |R|}{|C|} \tag{7}$$

式中: $\gamma_R(D)$ 为在特征子集 R 上训练出的分类器的分类精度; $|R|$ 为选择出的恶意软件特征子集包含的特征的数目; $|C|$ 为原始特征集合中特征总数; α 和 β 分别为用来调节分类精度和特征数目比重的参数, α 的取值为 $(0, 1)$, $\beta = 1 - \alpha$ 。在实验中, α 取值为 0.8, β 取值为 0.2。

其次,特征将按照式(8)进行更新:

$$X_i = \begin{cases} 1 & r < \frac{1}{1 + e^{-v_i}} \\ 0 & r \geq \frac{1}{1 + e^{-v_i}} \end{cases} \tag{8}$$

式中: r 为属于 $(0, 1)$ 区间的一个随机数; v_i 为粒子的速度大小。

最后,设置停止条件。当选择出的特征子集在分类器中的检测准确率达到阈值时则停止,否则继续进行特征选择。

2 实验分析

2.1 实验环境和条件

实验所用软件资源如表 3 所示,包括 Python

和 Anaconda,分别用来提供脚本语言环境和搭建虚拟环境。实验所用硬件资源如表 4 所示,使用 Mac 计算机来编写代码和运行程序。

表 3 实验所用软件资源

Table 3 Software resources used in experiment	
软件名称	来源
Python 3.7.3	https://www.python.org/ (开源)
Anaconda 4.3.3	https://www.anaconda.com/ (开源)

表 4 实验所用硬件资源

Table 4 Hardware resources used in experiment	
名称	描述
实验计算机	Mac
操作系统	MacOS High Sierra
网络配置	3.1 GHz Intel Core i5 处理器,8 GB 2 133 MHz 内存

2.2 评价方法

将准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 和 F_1 值作为实验评价指标,计算公式如下:

Precision = $\frac{tp}{tp + fp}$

(9)

Recall = $\frac{tp}{tp + fn}$

(10)

$F_1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$

(11)

Accuracy = $\frac{tp + tn}{tp + tn + fp + fn}$

(12)

式中:tp 为恶意软件被正确检测的样本数量;tn 为良性软件被正确检测的样本数量;fp 为良性软件被检测为恶意软件的样本数量;fn 为恶意软件被检测为良性软件的样本数量。

2.3 检测能力评估实验

2.3.1 实验目的和数据源

本节实验旨在评估本文方法在不同算法上使用前后的检测性能表现。AMD 和 DREBIN 公开数据集为实验数据源,数据集概况如表 5 所示。

表 5 数据集概况

Table 5 Description of dataset				
数据集	软件总数量	恶意软件数量	良性软件数量	类别
DREBIN	11 120	5 560	5 560	2
AMD	49 300	24 650	24 650	2

表 5 中,AMD 数据集包含了 2010—2016 年间的 24 650 个恶意软件样本及 24 650 个良性样本,恶意软件包含了 135 个恶意软件类型和 71 个恶意软件家族。DREBIN 数据集包含了 2010 年 8 月至 2012 年 10 月的 5 560 个恶意软件和 5 560 个良性软件。由于广告软件处于恶意软件

和良性功能之间的模糊地带,该数据集中不含有广告软件样本。

2.3.2 实验过程和参数

首先,将 DREBIN 数据集集中的 80% 数据作为训练集,20% 的数据作为测试集。

然后,利用训练集选择出模型中合适的参数。其中,Mann-Whitney 检验拒绝零假设的阈值 α 为 0.05,外观比率比例的阈值 $T_{\text{proportion}}$ 取值为 0.2,外观比率比例差异的阈值 $T_{\text{difference}}$ 取值为 0.5,粒子群优化算法中 α 取值为 0.8, β 取值为 0.2,加速因子 c_1 为 0.5, c_2 为 0.5。

其次,在训练集下,分别利用原始特征集和最优特征集训练梯度提升树算法 (GBDT)^[12]、自适应提升算法 (AdaBoost)^[13]、多层神经网络算法 (MLP)^[14]、逻辑回归算法 (LR)^[15]、贝叶斯算法 (NB)^[16] 和随机森林^[17] 算法。其中,随机森林算法中树的数量为 100,每个决策树随机选择的特征数目为 241。

最后,在测试集下,分别利用原始特征集和最优特征子集计算上述算法的准确率、精确率、召回率和 F_1 值,最终得到 Android 恶意软件检测外观比率间隔特征选择方法的实验结果。

AMD 数据集重复以上过程。

2.3.3 实验结果

以 AMD 数据集为例,使用提出的外观比率间隔特征选择方法选择出维度为 294 的最优特征子集,如表 6 所示。检测能力评估实验的结果如表 7 和表 8 所示。

表 6 AMD 数据集的最优特征子集

Table 6 Optimal feature subset of AMD dataset		
特征类型	数目	特征 (AMD 数据集)
系统权限	14	ACCESS_NETWORK_STATE
		ACCESS_WIFI_STATE
		BROADCAST_STICKY
		CAMERA
		GET_TASKS
API	262	
		android/net/wifi/WifiManager;enableNetwork
		android/net/wifi/WifiManager;setWifiEnabled
		Dex Class Loader
		
组间通信	18	android.intent.action.ACTION_SHUTDOWN
		android.intent.action.AIRPLANE_MODE
		android.intent.action.BOOT_COMPLETED
		android.intent.action.MEDIA_MOUNTED
		android.intent.action.SEARCH
Intent	18	

表 7 特征选择实验结果 (AMD 数据集)

Table 7 Experimental results of feature selection (AMD dataset)

方法	原始特征集 (689 维)				最优特征集 (294 维)			
	Accuracy/%	Precision/%	Recall/%	F_1	Accuracy/%	Precision/%	Recall/%	F_1
GBDT	95.0	94.9	95.2	0.950	95.6	96.2	95.1	0.956
MLP	96.0	96.1	96.1	0.961	96.4	96.0	97.4	0.970
LR	92.9	90.2	95.5	0.927	94.6	95.2	95.2	0.947
AdaBoost	93.2	92.3	94.2	0.932	93.8	94.3	93.5	0.939
NB	86.4	85.3	87.5	0.864	85.0	94.5	88.9	0.864
RF	97.6	96.7	98.5	0.974	98.2	98.1	98.4	0.983

表 8 特征选择实验结果 (DREBIN 数据集)

Table 8 Experimental results of feature selection (DREBIN dataset)

方法	原始特征集 (689 维)				最优特征集 (295 维)			
	Accuracy/%	Precision/%	Recall/%	F_1	Accuracy/%	Precision/%	Recall/%	F_1
GBDT	96.8	90.4	94.3	0.924	96.9	92.0	94.5	0.935
MLP	97.0	93.3	92.7	0.930	98.4	95.0	94.0	0.944
LR	96.3	88.8	93.3	0.924	96.1	89.0	93.1	0.924
AdaBoost	96.0	90.0	91.3	0.903	96.8	89.9	92.2	0.916
NB	94.6	82.1	91.4	0.865	95.3	86.7	91.9	0.871
RF	98.1	93.2	98.2	0.956	98.9	94.9	98.9	0.967

2.4 对比分析实验

2.4.1 实验目的和数据源

本节实验旨在对比分析提出的外观比率间隔特征选择方法与其他特征选择方法在 Android 恶意软件检测准确率上的性能差异。DREBIN 公开数据集为实验数据源,其详细介绍如表 5 所示。

2.4.2 实验过程和参数

首先,将 DREBIN 数据集中的 80% 数据作为训练集,20% 的数据作为测试集。

然后,利用训练集选择出模型中合适的参数。其中,Mann-Whitney 检验拒绝零假设的阈值 α 为 0.05,外观比率比例的阈值 $T_{\text{proportion}}$ 取值为 0.5,外观比率比例差异的阈值 $T_{\text{difference}}$ 取值为 0.5,粒子群优化算法中 α 取值为 0.8, β 取值为 0.2,加速因子 c_1 为 0.5, c_2 为 0.5。

其次,在训练集下,分别利用文献[18-22]中提到的卡方检验、方差选择、递归特征消除、遗传算法和关联规则特征选择方法,与本文方法对原始特征集进行特征选择并训练随机森林分类器。其中,随机森林算法中树的数量为 100,每个决策树随机选择的特征数目为 241。

最后,在测试集下,分别利用文献[18-22]中的方法,与本文方法选择出特征并测试随机森林分类器的准确率、精确率、召回率和 F_1 值,最终得到 Android 恶意软件检测外观比率间隔特征选择方法对比实验结果。

2.4.3 实验结果

对比分析实验结果如表 9 所示。

表 9 特征选择方法实验结果对比

Table 9 Comparison of experimental results among feature selection methods

方法	特征数量	Accuracy/%	Precision/%	Recall/%	F_1
文献[18]	364	96.5	96.0	97.3	0.967
文献[19]	394	96.4	96.4	96.9	0.967
文献[20]	400	97.0	96.8	97.3	0.971
文献[21]	314	96.4	96.4	96.8	0.966
文献[22]	22	96.1	97.0	97.4	0.962
本文方法	294	98.2	98.1	98.4	0.983

3 实验结果讨论

在生成最优特征子集时,外观比率间隔特征选择方法可以挑选出低频和低偏差的特征,同时将不相关特征及冗余特征从备选特征集中剔除。从表 6 可以看出,Android 恶意软件更倾向于使用与 SMS 短信相关的特征,这类特征在正常应用中所使用的频率较低,同时 Dex Class Loader 等在 Android 正常应用中所使用的频率远高于在恶意软件中所使用的频率。这一现象表明,Android 开发人员有意识地通过动态加载这一技术对自己的核心代码进行保护,同时开发人员还可以通过该技术对自己的软件进行动态拓展。

在检测能力评估实验中,本文方法能够提升分类方法的检测效果。将本文方法应用到分类方法后显示(见表 7),在 AMD 数据集中部分基础分类方法的检测效果有所提升,如使用 GBDT 方法进行恶意软件检测,其精确率从 94.9% 提升至

96.2%, LR 方法的精确率从 90.2% 提升至 95.2%, MLP 方法的召回率和 F_1 值均提升了 0.01, AdaBoost 方法、NB 方法等准确率、精确率都有不同程度的提升。表 8 显示,在 DREBIN 数据集中,部分基础分类方法的检测效果有所提升,如使用 RF 方法进行 Android 恶意软件检测,其精确率从 93.2% 提升至 94.9%,其他方法的准确率、精确率等评价指标在进行特征选择后都有不同程度的提升。本文方法能够拉近同类特征在特征空间中的距离,不易过拟合,能够准确检测 Android 恶意软件。因此,本文方法能够提升分类算法的检测效果。

在对比分析实验中,本文方法优于其他对比方法。结果显示,该方法选择出的特征数量为 294,准确率为 98.2%,召回率即检测率为 98.4%,精确率为 98.1%, F_1 值为 0.983,特征数量少于 4 种(共 5 种)对比方法,准确率、精确率、召回率和 F_1 值均优于其他 5 种对比方法。该方法通过分离类间存在频率分布偏差的特征,量化偏差程度和特征出现频率以剔除低偏差和低频特征,并结合分类器实际分类效果选择出特征子集,相比于其他利用特征与检测结果相关关系的对比方法(卡方检验、方差选择、递归特征消除、遗传算法和关联规则),具有更好的特征选择效果。因此,本文方法优于其他对比方法。

4 结 论

本文针对现有的 Android 恶意软件检测特征选择方法忽略特征在类间的频率分布情况,对同时出现在恶意软件和良性软件的特征过度关注,导致特征冗余性较强的问题,提出一种 Android 恶意软件检测外观比率间隔特征选择方法。

1) 该方法利用 Mann-Whitney 检验方法选择出在恶意软件和良性软件中存在频率分布偏差的特征,通过外观比率间隔算法量化偏差程度和特征出现频率剔除低偏差和整体软件中低频使用的特征,利用粒子群优化算法根据实际检测效果得到最优特征子集,降低了 Android 恶意软件特征集的冗余性。

2) 使用 DREBIN 公开数据集和 AMD 公开数据集进行特征选择能力评估实验和对比实验,实验结果显示,在 AMD 数据集上选择出了 294 维特征,进行特征选择后 6 种分类器的检测准确率提高了 1%~5%,最佳检测准确率为 98.2%;在 DREBIN 数据集上选择出了 295 维特征,进行特征选择后将随机森林分类器的检测精确率提高了

1.7%。与 5 种特征选择方法对比,具有较少的特征数量,具有较高的准确率、精确率、召回率(检测率)和 F_1 值。这是因为本文方法可以通过确定阈值保存其他方法丢弃的有用特征,同时去除其他方法的冗余特征,降低了特征的冗余性,提高了恶意软件检测特征选择方法的检测效果。

研究工作还可以进一步增加从 Apk 文件中提取出的初始特征类型和数量,从而提升 Android 恶意软件检测能力,实现对 Android 系统和用户进行全面的安全防护。

参考文献 (References)

[1] 中国互联网络信息中心.第 44 次中国互联网络发展现状统计报告[R].北京:中国互联网络信息中心,2019.
China Internet Network Information Center. The 44th China statistical reports on internet development[R]. Beijing: China Internet Network Information Center,2019 (in Chinese).

[2] International Data Corporation. Worldwide smartphone market shares[R]. New York:International Data Corporation,2019.

[3] YERIMA S Y,SEZER S,MCWILLIAMS G. Analysis of Bayesian classification-based approaches for Android malware detection[J]. IET Information Security,2014,8(1):25-26.

[4] PEHLIVAN U,BALTACI N,ACARTURK C,et al. The analysis of feature selection methods and classification algorithms in permission based Android malware detection[C]//Computational Intelligence in Cyber Security. Piscataway: IEEE Press,2014:1-8.

[5] WANG W,WANG X,FENG D,et al. Exploring permission-induced risk in Android applications for malicious application detection[J]. IEEE Transactions on Information Forensics and Security,2014,9(11):1869-1882.

[6] CEN L,GATES C S,SI L,el al. A probabilistic discriminative model for Android malware detection with decompiled source code[J]. IEEE Transactions on Dependable and Secure Computing,2015,12(4):400-412.

[7] ZHAO K,ZHANG D,SU X,et al. Fest: A feature extraction and selection tool for Android malware detection[C]//Computers and Communication. Piscataway: IEEE Press,2015:714-720.

[8] TAO G,ZHENG Z,GUO Z,et al. MalPat: Mining patterns of malicious and benign Android apps via permission-related APIs [J]. IEEE Transactions on Reliability,2018,67(1):355-369.

[9] LI J,SUN L,YAN Q,et al. Significant permission identification for machine learning based Android malware detection [J]. IEEE Transactions on Industrial Informatics,2017,14(7):3216-3225.

[10] DESNOS A,GUEGUEN G,BACHMANN S. Androguard package [EB/OL]. (2020-04-30) [2021-09-01]. <https://github.com/androguard>.

[11] MANN H B,WHITNEY D R. On a test whether one of two random variables is statistically larger than the other[J]. Annals of Mathematical Statistics,1947,18(1):50-60.

[12] FRIEDMAN J H. Stochastic gradient boosting[J]. Computation-

al Statistics & Data Analysis,2002,38(4):367-378.

[13] FREUND Y,SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences,1997,55(1):119-139.

[14] HANSEN L K,SALAMON P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(10):993-1001.

[15] RUCZINSKI I,KOOPERBERG C,LEBLANC M. Logic regression[J]. Journal of Computational and Graphical Statistics, 2003,12(3):475-511.

[16] FRIEDMAN N,GEIGER D,GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning,1997,29:131-163.

[17] BREIMAN L. Random forests[J]. Machine Learning,2001,45: 5-32.

[18] MORALES O S,ESCAMILLA A P J,RODRGUEZ M A,et al. Native malware detection in smartphones with Android OS using static analysis,feature selection and ensemble classifiers[C]// International Conference on Malicious and Unwanted Software. Piscataway:IEEE Press,2016:67-74.

[19] SEDANO J,GONZLEZ S,CHIRA C,et al. Key features for the characterization of Android malware families[J]. Logic Journal of the IGPL,2017,25(1):54-66.

[20] RAI S,DHANESHA R,NAHATA S,et al. Malicious application detection on Android smartphones with enhanced static-dynamic analysis[C]// International Conference on Information Systems Security. Berlin:Springer,2017:194-208.

[21] FATIMA A,MAURYA R,DUTTA M K,et al. Android malware detection using genetic algorithm based optimized feature selection and machine learning[C]// International Conference on Telecommunications & Signal Processing. Piscataway: IEEE Press,2019:220-223.

[22] SUN L,LI Z,YAN Q,et al. SigPID:Significant permission identification for android malware detection[C]// International Conference on Malicious and Unwanted Software. Piscataway: IEEE Press,2017:1-8.

Low redundancy feature selection method for Android malware detection

HAO Jingwei¹, PAN Limin¹, LI Rui¹, YANG Peng^{2,*}, LUO Senlin¹

(1. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China;

2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: A low redundancy feature selection method for Android malware detection is proposed to solve the problem of feature redundancy caused by excessive attention to features with the same frequency distribution between classes. First, the method selects features with frequency distribution bias by Mann-Whitney test, and then quantifies the degree of bias and feature appearance frequency by the appearance ratio interval algorithm to reject features with low bias and low use frequency in the overall software. Finally, the particle swarm optimization algorithm is combined with model detection effect to obtain the optimal feature subset. Experiments were conducted using public datasets DREBIN and AMD. The experimental results show that 294-dimensional features were selected on the AMD dataset, and the detection accuracy of the six classifiers is improved by 1% – 5%, 295-dimensional features were selected on the DREBIN dataset less than 4 comparison methods, and the detection accuracy of the six classifiers is improved by 1.7% – 5%. The experimental results illustrate that the proposed method can reduce the redundancy of features in Android malware detection and improve the malware detection accuracy.

Keywords: Android malware detection; feature selection; Mann-Whitney test; particle swarm optimization algorithm; appearance ratio interval algorithm

Received: 2020-09-30; **Accepted:** 2020-12-18; **Published online:** 2021-03-02 14:23
URL: kns.cnki.net/kcms/detail/11.2625.V.20210302.0900.001.html
Foundation items: 242 National Information Security Projects (2019A012); 2020 Information Security Software Project of the Ministry of Industry and Information Technology (CEIEC-2020-ZM02-0134)
*** Corresponding author.** E-mail: yp@cert.org.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0383

基于随机森林的物联网设备流量分类算法

李锐光^{1,2}, 段鹏宇¹, 沈蒙¹, 祝烈煌^{1,*}

(1. 北京理工大学 网络空间安全学院, 北京 100081; 2. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 物联网 (IoT) 设备流量分类对网络资产管理有重要意义, 基于流量统计的分类技术是当前研究热点。已有算法主要基于流信息建立特征向量, 而对数据包信息利用较少。改进了基于随机森林的物联网设备流量分类算法, 基于流信息和流数据包信息共同建立特征向量。实验结果表明: 所提算法与其他算法相比, 所提算法的平均分类准确率由 56% 提高到 82%, 平均召回率由 47% 提高到 67%, 平均 F_1 得分由 0.43 提高到 0.74, 混淆矩阵对比也有明显提升, 因此具备更好的分类效果。

关 键 词: 物联网 (IoT); 流量分类算法; 随机森林; 特征向量; 流信息; 数据包信息
中图分类号: TP393; TN919

文献标志码: A **文章编号:** 1001-5965(2022)02-0233-07

物联网技术将各种带有传感器的终端设备接入互联网, 组成网络并实现人和物的互联互通。目前, 物联网技术在家庭、校园、企业、城市等应用场景已得到广泛应用^[1], 海量的网络摄像头、身体传感器、智能家电等物联网设备接入互联网, 给人类生活带来了重大变革。

然而, 在物联网广泛普及的同时, 政府监管机构、网络运营者对网络空间中的物联网设备缺少统计和管理。一方面, 各种类型的物联网设备通常由不同运营者、不同厂商、不同人员安装, 对全量设备的登记和统计非常困难; 另一方面, 物联网广泛采用了 NAT 技术和 IP 地址动态分配, 缺少唯一化的标记手段, 使得掌握网络中物联网设备总体情况变得困难。

由于物联网设备普遍资源受限导致的安全防护不足, 使得很多物联网设备成为网络攻击的工具或跳板, 对全球网络空间安全带来严峻挑战。因此, 对网络空间中的海量物联网设备的识别和分类, 成为当前工业界和学术界关注的热点研究问题。

现有的物联网设备流量分类算法主要有基于

协议特征识别和基于流量统计两大类。基于协议特征识别需要找到流量强特征, 在目前普遍加密的网络中已越来越难, 因此基于流量统计的识别方法逐步成为研究热点。本文主要对当前流量识别领域广泛应用的随机森林算法进行了改进, 并取得了不错的实验效果。

1 物联网设备识别方法

现有的物联网设备识别方法主要包括主动资源探测和被动流量分析两大类。

1) 主动资源探测是按照物联网设备通信协议格式, 向远端设备发送探测包, 采集和分析响应数据, 从而得到远端设备重要信息的识别方法。通过主动探测, 可以得到操作系统、开放端口、漏洞分布、设备类型、厂商型号、所属组织等重要信息, 从而掌握网络中全量物联网设备情况。Feng 等^[2]使用 Modbus 和 S7 协议标语的格式和信息识别工控设备的厂商和设备版本。Li 等^[3-4]使用设备的登录页面提取关键字从而识别设备, 并测量页面中单词的集中度, 将搜索引擎返回的结果

收稿日期: 2020-08-03; 录用日期: 2020-09-05; 网络出版时间: 2020-10-20 09:36
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201019.1459.002.html
* 通信作者。E-mail: liehuangz@bit.edu.cn

引用格式: 李锐光, 段鹏宇, 沈蒙, 等. 基于随机森林的物联网设备流量分类算法[J]. 北京航空航天大学学报, 2022, 48(2): 233-239. LI R G, DUAN P Y, SHEN M, et al. Traffic classification algorithm of Internet of things devices based on random forest[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 233-239 (in Chinese).

作为设备指纹。Leonard 和 Loguinov^[5] 使用 TCP ACK 探测主机是否存活,使用 DNS、HTTP、SMTP 等协议的响应识别主机的设备类型和操作系统。

2) 被动流量分析是指从某一网络关口采集流量数据,分析和提取不同类型物联网设备的协议指纹或流量特征,从而实现对物联网设备类型的识别。Kohno 等^[6] 使用多种协议的时钟偏移,即时钟时间与实际时间的差进行设备识别。Aneja 等^[7] 使用到达间隔时间 (IAT) 图作为特征使用 CNN 算法进行分类。Husák 等^[8] 使用 HTTPS 协议中的 SSL/TLS 握手信息作为设备指纹对用户代理进行分类。

主动资源探测方法通常用于对某一网域以外设备的探测识别,而被动流量分析通常用于对某一网域内设备的识别分类。本文重点研究被动流量分析方法,提出并实现了一种基于随机森林的物联网设备流量分类算法。

2 物联网设备流量分类技术

物联网设备流量分类技术分为 2 类:①协议特征识别,即依靠协议中存在的强特征对不同设备流量进行匹配分类,如文献[2,8];②统计识别,即寻找流量数据中的包数、包长、时间间隔、请求数量等弱特征,采用统计的方法对不同设备流量进行分类识别,如文献[6-7]。

协议特征识别方法的优点是识别准确率高,但需要提前知道设备类型及其通信协议,对于未知的物联网设备无法识别,而且随着加密技术在网络协议传输中的广泛采用,协议特征变得越来越模糊,因此,基于协议强特征的识别方法越来越表现出局限性。基于统计的识别方法有可能识别出未知的物联网设备,因为同一厂家产品在流量上具有相似性,另外可以有效地识别和分类加密流量。

学术界对基于统计的设备流量分类算法开展了大量研究。Arunan 等^[9] 使用流的信息和 DNS、NTP 的请求数量、请求间隔等进行分类。Msadek 等^[10] 将端口号分为公认端口、注册端口、动态端口,结合协议类型和流统计信息对设备流量进行分类。Yao 等^[11] 提出一种可以自动选择流量特征、基于深度学习的物联网设备分类算法,主要应用于智慧城市市场。Desai 等^[12] 仅使用流的统计信息,包括均值、中位数、方差、分布相似度进行分类。Meidan 等^[13] 除了使用流信息,还使用其他数据,包括 Alexa 排名和 GeoIP 信息进行分类。Shahid 等^[14] 使用发送和接收到的前 N 个包的大小、IAT 进行分类。

在基于统计的流量分类技术中,特征向量(统计参数)的选择至关重要,直接影响分类准确

率及算法性能。大多数研究中采用了完整的 TCP/UDP 流的信息,有的辅助采用了 DNS、NTP 的请求数量、请求间隔^[9],有的辅助采用了域名、地理位置^[13] 等参数信息。本文在现有工作基础上,重点研究了 TCP/UDP 流的数据包信息,在特征向量中加入包长、包数、包 TLL 等参数,实验表明提升了分类准确性。

3 数据集

本文在研究中采用了 Sivanathan 等^[15] 公开的数据集。该数据集来自一所学校网关处捕获的真实流量,持续时间为 2016 年 9 月 23 日至 10 月 6 日。实验者在网关处部署了 tcpdump 软件捕获流量,使用 kmod-usb-core 和 kmod-usb-storage 软件存储流量数据。

数据集中包括了 21 种物联网设备和 9 种非物联网设备的混合流量,以 PCAP 包格式按天保存,总大小为 9 GB。由于数据集中包含了各设备的 MAC 地址,这为从混合数据中提取各设备流量、建立训练数据集提供了很大便利。所有设备名称及类型如表 1 所示。

表 1 物联网设备
Table 1 IoT device list

编号	设备名称	设备类型
1	Smart Things	物联网设备
2	Amazon Echo	物联网设备
3	Netatmo Welcome	物联网设备
4	TP-Link Day Night Cloud camera	物联网设备
5	Samsung SmartCam	物联网设备
6	Dropcam	物联网设备
7	Insteon Camera 1	物联网设备
8	Insteon Camera 2	物联网设备
9	Withings Smart Baby Monitor	物联网设备
10	Belkin Wemo switch	物联网设备
11	TP-Link Smart plug	物联网设备
12	iHome	物联网设备
13	Belkin wemo motion sensor	物联网设备
14	NEST Protect smoke alarm	物联网设备
15	Netatmo weather station	物联网设备
16	Withings Smart scale	物联网设备
17	Blipcare Blood Pressure meter	物联网设备
18	Withings Aura smart sleep sensor	物联网设备
19	LiFX Smart Bulb	物联网设备
20	Triby Speaker	物联网设备
21	PIX-STAR Photo-frame	物联网设备
22	HP Printer	非物联网设备
23	Samsung Galaxy Tab	非物联网设备
24	Nest Dropcam	非物联网设备
25	Android Phone	非物联网设备
26	Laptop	非物联网设备
27	MacBook	非物联网设备
28	Android Phone	非物联网设备
29	IPhone	非物联网设备
30	MacBook/Iphone	非物联网设备

数据集中的物联网设备类型包括网络摄像头、智能开关、空气质量传感器、智能音箱、智能灯泡、智能相框及医疗保健设备等。非物联网设备包括打印机、笔记本电脑、智能手机、平板电脑等。设备厂商涵盖了亚马逊、TP-Link、三星、贝尔金、Netatmo、苹果、惠普、NEST、谷歌等主流设备厂商。因此,基于该数据集开展的研究工作具有很好的普适性。

4 算法设计

本文改进了文献[15]提出的基于随机森林的物联网设备流量分类算法,在特征选择上除了流信息外也将每条流数据包信息作为统计参数,算法过程如下。

4.1 数据预处理

文献[15]公开的数据集中保存了从网关捕获的全部设备的混合流量,为了对机器学习模型进行训练和测试,先需要对流量做预处理,获取训练集 D_T 和测试集 D_M 。

原始流量数据是按天保存的,文献[15]直接使用设备一天的流量作为样本数据。由于设备一天的流量数据较多,在训练模型时耗费的时间较长。采用设备连接后几分钟内的流量数据作为样本,减少了机器学习模型训练时间,实验表明分类效果未受影响。

首先,对全部流量按 MAC 地址进行分割,然后,按算法 1 筛选出每类设备建立连接后几分钟内的流量,作为样本数据保存,最终,获取到 80 万条样本数据。对全部样本数据按设备类型进行了分类标注。

算法 1 流量分割算法。

步骤 1 输入时间阈值 t_T 。

步骤 2 读取未保存的第一个流, t_{s1} 为流开始时间, t_{e1} 为流结束时间。

步骤 3 读取下一个流, t_{s2} 为流开始时间,如果 $t_{s2} - t_{e1} > t_T$, 将从 $t_{s1} \sim t_{s1} + t_T$ 之间的所有数据包保存为一条数据,返回步骤 1; 如果 $t_{s2} - t_{e1} < t_T$, 流结束时间大于 t_{e1} , 将 t_{e1} 保存为新的流的结束时间,重复步骤 3。

在这 80 万条数据中包含了表 1 中每类设备的样本数据,平均每类设备有 2 万 ~ 3 万条样本。从各类设备样本中选取 90% 作为训练集 D_T , 剩余样本混合后作为测试集 D_M 。

4.2 特征提取

在处理后的样本数据中,编制脚本程序从中自动提取出流信息和数据包信息。流信息包括流

数量、协议种类、最常用协议、端口数量、最常用端口等。数据包信息包括发送和接收到的总字节数和总包数信息、发送和接收数据包的 TTL、发出和接收数据包的首个时间窗口等。

从样本中自动提取特征的算法如算法 2 所示。

算法 2 特征提取算法。

步骤 1 选取下一条流。

步骤 2 记录流的个数、采用的协议、连接的端口。

步骤 3 选取流中的下一个包。

步骤 4 如果包是流中第一个包,记录接收窗口。

步骤 5 记录包的长度、包的个数、数据包 TTL。

步骤 6 如果包是流中最后一个包,返回步骤 1, 否则返回步骤 3。

步骤 7 统计这条样本中流的数量、协议种类、最常用的协议、连接端口数量、最常用的端口、发送和接收的总字节数和总包数。

本文算法的特征向量包括流信息和包信息,如表 2 所示。

表 2 特征向量
Table 2 Feature vectors

名称	释义	类型
流数量	每条样本中包含流的数量	流信息
协议种类	每条样本中包含的协议种类	
最常用协议	每条样本中包含数量最多的协议	
端口数量	每条样本中连接过的端口数量	
最常用端口	每条样本中连接次数最多的端口	
发送总字节数	每条样本中发送总字节数	包信息
接收总字节数	每条样本中接收总字节数	
发送总包数	每条样本中发送总包数	
接收总包数	每条样本中接收总包数	
发送包 TTL	每条样本中首个发送数据包存活时间	
接收包 TTL	每条样本中首个接收数据包存活时间	
首包发送窗口	每条样本中首条流的数据包发送窗口	
首包接收窗口	每条样本中首条流的数据包接收窗口	

4.3 算法实现

在分类算法设计上,近年来机器学习算法得到了广泛应用并取得良好效果,如朴素贝叶斯 (naive Bayes)、支持向量机 (supporting vector machine)、K 近邻 (KNN)、决策树 (decision trees)、随机森林 (random forest)、神经网络 (neural network) 等。通常,流量分类场景中特征向量复杂度远少于图像处理、自然语言处理等场景,因此一般不必采用太复杂的机器学习算法 (如深度学习) 即可取得良好效果。

随机森林是一种使用多棵决策树进行训练和

预测的分类器,从原始的训练数据中有放回地抽取 k 个样本构成 m 个训练集,在每个训练集上构造一棵决策树。构造决策树时,在节点寻找特征进行分裂时,随机选取一些特征,在抽取到的特征中选择最优解进行分裂,最终根据多数投票原则产生分类结果。该算法已在流量分类等应用场景中得到了广泛应用。

根据表 2 中的特征向量,在 Python 集成开发环境中利用 Scikit-Learn 工具包实现了随机森林算法。设置时间阈值 t_T 为 10 min,设置随机森林算法中树的数量为 10。利用每类设备的训练集 D_T 对模型进行训练,在测试集 D_M 上进行了测试,并统计了分类准确率、召回率、 F_1 得分等情况。

5 实验结果

基于公开数据集,对文献[15]算法和本文算法进行了对比测试,并比较了分类准确率、召回率、 F_1 得分、混淆矩阵等指标。

5.1 评价标准

对机器学习算法的评价标准有分类准确率、召回率和 F_1 得分,也包括混淆矩阵。例如,预测为正的样本中,实际为正的样本数为 TP,实际为负的样本数为 FP;预测为负的样本中,实际为正的样本数为 FN,实际为负的样本数为 TN。

分类准确率定义为

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

召回率定义为

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

F_1 得分定义为

$$F_1 = \frac{2\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

混淆矩阵也是衡量分类算法性能的一种方式,如图 1 所示。矩阵中第 i 行第 k 列表示将第 i 种设备错误识别成第 k 种设备的样本数量,颜色越深表示样本数越多。这样,如果对角线位置的方块颜色越深而其他位置颜色越浅,则表示分类器分类的准确性越高。

5.2 文献[15]中算法分类效果

在公开数据集上复现了文献[15]中提出的算法,并进行了测试。该算法以某设备一天内的流量数据作为样本集,在特征选择上只采用样本集中流信息作为统计参数。该算法分类效果如表 3 所示。

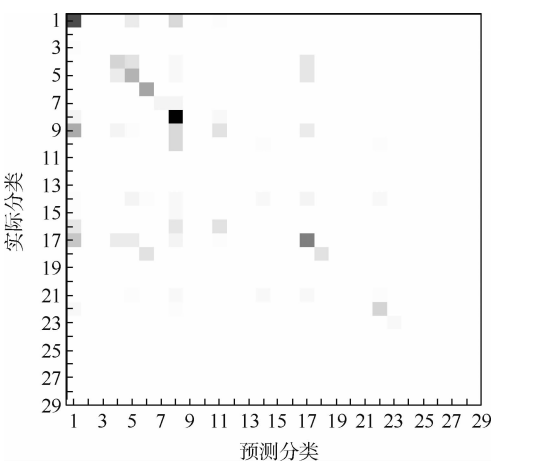


图 1 文献[15]中算法的混淆矩阵
Fig. 1 Confusion matrix of algorithm in Ref. [15]

表 3 文献[15]中算法的分类效果
Table 3 Classification results of algorithm in Ref. [15]

分类	准确率/%	召回率/%	F_1 得分
Amazon Echo	44	69	0.54
Android Phone 1	0	0	0
Android Phone 2	0	0	0
Belkin Wemo switch	42	37	0.39
Belkin wemo motion sensor	43	51	0.47
Dropecam	66	100	0.79
HP Printer	98	44	0.61
Insteon Camera 1	58	94	0.72
Laptop	0	0	0
LiFX Smart Bulb	0	0	0
MacBook	1	12	0.01
NEST Protect smoke alarm	0	0	0
Nest Dropecam	0	0	0
Netatmo Welcome	24	12	0.16
Netatmo weather station	100	5	0.09
Samsung Galaxy Tab	0	0	0
Samsung SmartCam	55	50	0.52
Smart Things	91	47	0.62
TP-Link Day Night Cloud camera	0	0	0
TP-Link Smart plug	0	0	0
Triby Speaker	65	0	0.01
Withings Aura smart sleep sensor	62	68	0.65
Withings Smart Baby Monitor	33	74	0.45

文献[15]中算法形成的混淆矩阵如图 1 所示,图中横竖坐标数字对应表 1 中的设备序号。可见,除对角线外,还零散分布着一些样本点。

5.3 本文算法分类效果

本文对文献[15]中算法进行了改进,一方面是将某设备连接建立后几分钟内的数据作为样本,而不是全天数据,另一方面是除了流信息外加入了数据包信息作为统计参数。本文算法的分类效果如表 4 所示。

本文算法的混淆矩阵如图 2 所示,图中横竖坐标数字对应表 1 中的设备序号。与图 1 比较,样本点基本集中在对角线处,除对角线外零散分布的样本点很少,因此分类效果优于图 1。

表 4 本文算法的分类效果

Table 4 Classification results of proposed algorithm

分类	准确率/%	召回率/%	F_1 得分
Amazon Echo	69	84	0.76
Belkin Wemo switch	60	58	0.59
Belkin wemo motion sensor	63	29	0.4
Dropcam	100	99	1
HP Printer	86	66	0.75
Insteon Camera 1	87	99	0.92
LiFX Smart Bulb	97	41	0.57
NEST Protect smoke alarm	0	0	0
Nest Dropcam	0	0	0
Netatmo Welcome	64	61	0.62
Netatmo weather station	98	86	0.91
PIX-STAR Photo-frame	0	0	0
Samsung SmartCam	58	92	0.71
Smart Things	97	86	0.91
TP-Link Day Night Cloud camera	0	0	0
TP-Link Smart plug	0	0	0
Triby Speaker	89	43	0.59
Withings Aura smart sleep sensor	91	72	0.81
Withings Smart Baby Monitor	60	91	0.73
Withings Smart scale	0	0	0
iHome	100	9	0.16
non-IoT	89	53	0.66

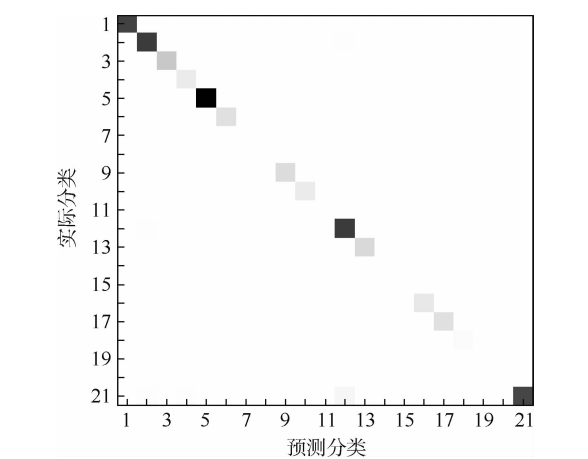


图 2 本文算法的混淆矩阵

Fig.2 Confusion matrix of proposed algorithm

5.4 对比分析

通过 5.2 节和 5.3 节的实验结果表明:

1) 本文算法对真实流量分类的平均分类准确率达 82%,平均召回率达 67%,平均 F_1 得分为 0.74,具有较好的实用性。

2) 文献[15]中算法的平均分类准确率为 56%,平均召回率为 47%,平均 F_1 得分为 0.43,可见本文算法比文献[15]中算法分类性能有明

显提升。

3) 在设备流量分类技术的特征选择上,单纯选择流信息的分类效果并不很好,而加入数据包信息后的分类效果提升明显。

4) 表 3 中 Android Phone、Laptop、LiFX Smart Bulb 等设备和表 4 中的 NEST Protect smoke alarm、Netatmo Welcome 等设备未测到分类准确率、召回率和 F_1 得分数据,原因是测试集 D_M 中数据是从 80 万条样本中随机选择的,未包括这些设备流量。

5) 本文算法对 Belkin Wemo switch 和 Belkin wemo motion sensor 两个设备分类效果较差,可能是由于同一厂商的设备行为比较相似,在流量上难以分辨。

6) 2 种算法的运算时间和内存占用情况如表 5 所示。由于本文算法特征向量维度大于文献[15]算法,故运算时间及内存占用均略高,但仍处于可接受范围之内。

本文算法与文献[15]算法相比,在统计参数的选择上更加精确,因此具有更好的分类效果。但由于特征向量维度的增加,本文算法实际消耗的计算和存储资源略有增长,因此更适用于对物联网设备的精确识别场景。

表 5 算法性能对比

Table 5 Algorithm performance comparison

算法	运算时间/s	内存/MB
文献[15]算法	14.5	342
本文算法	18	451

6 结 论

本文提出一种基于随机森林的物联网设备流量分类算法。在文献[15]公开数据集和提出算法的基础上,本文对文献[15]算法做了改进,主要如下:

- 1) 将设备连接建立后几分钟内的数据作为样本,而没有采用全天数据,从而缩短了模型训练时间。
- 2) 在特征选择中除了流信息外加入了完整流的数据包信息。
- 实验结果表明,本文提出算法的平均分类准确性由 56% 提高到 82%,平均召回率由 47% 提高到 67%,平均 F_1 得分由 0.43 提高到 0.74,混淆矩阵对比也有明显提升,具备更好的分类效果。
- 由于数据集中设备数量限制,本文只对部分物联网设备进行了研究,分类识别也仅限于已知设备。未来应研究更多类型的物联网设备,并分

析归纳同厂家或同类设备的流量特点,探索对未知类型物联网设备的识别发现技术,更好地服务于网络空间安全管理。

参考文献 (References)

[1] 黄凯奇,陈晓棠,康运锋,等. 智能视频监控技术综述[J]. 计算机学报,2015,38(6):1093-1118.
HUANG K Q, CHEN X T, KANG Y F, et al. Intelligent visual surveillance: A review [J]. Chinese Journal of Computers, 2015, 38(6):1093-1118 (in Chinese).

[2] FENG X, LI Q, HAN Q, et al. Identification of visible industrial control devices at Internet scale[C] //2016 IEEE International Conference on Communications. Piscataway:IEEE Press, 2016:1-6.

[3] LI Q, FENG X, WANG H, et al. Automatically discovering surveillance devices in the cyberspace[C] // The 8th ACM. New York:ACM,2017:331-342.

[4] FENG X, LI Q, WANG H, et al. Acquisitional rule-based engine for discovering Internet-of-thing devices[C] //27th USENIX Security Symposium,2018:327-341.

[5] LEONARD D, LOGUINOV D. Demystifying service discovery: Implementing an Internet-wide scanner [C] // Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement 2010. New York:ACM,2010:109-122.

[6] KOHNO T, BROIDO A, CLAFFY K C. Remote physical device fingerprinting[J]. IEEE Transactions on Dependable and Secure Computing,2005,2(2):93-108.

[7] ANEJA S, ANEJA N, ISLAM M S. IoT device fingerprint using deep learning[C] //2018 IEEE International Conference on Internet of Things and Intelligence System. Piscataway: IEEE Press,2018:174-179.

[8] HUSÁK M, ČERMÁK M, JIRSÍK T, et al. HTTPS traffic analysis and client identification using passive SSL/TLS fingerprinting[J]. EURASIP Journal on Information Security,2016,2016(1):1-14.

[9] ARUNAN S, HASSAN H G, FRANCO L, et al. Classifying IoT devices in smart environments using network traffic characteristics[J]. IEEE Transactions on Mobile Computing,2019,18(8):1745-1759.

[10] MSADEK N, SOUA R, ENGEL T. IoT device fingerprinting: Machine learning based encrypted traffic analysis[C] // 2019 IEEE Wireless Communications and Networking Conference (WCNC). Piscataway:IEEE Press,2019:1-8.

[11] YAO H, GAO P, WANG J, et al. Capsule network assisted IoT traffic classification mechanism for smart cities[J]. IEEE Internet of Things Journal,2019,6(5):7515-7525.

[12] DESAI B A, DIVAKARAN D M, NEVAT I, et al. A feature-ranking framework for IoT device classification[C] // International Conference on Communication Systems & Networks, 2019:64-71.

[13] MEIDAN Y, BOHADANA M, SHABTAI A, et al. ProfilloT: A machine learning approach for IoT device identification based on network traffic analysis[C] //Proceedings of the Symposium on Applied Computing,2017:506-509.

[14] SHAHID M R, BLANC G, ZHANG Z, et al. IoT devices recognition through network traffic analysis[C] //IEEE International Conference on Big Data. Piscataway: IEEE Press, 2018:5187-5192.

[15] SIVANATHAN A, SHERRATT D, GHARAKHEILI H H, et al. Characterizing and classifying IoT traffic in smart cities and campuses [C] // IEEE INFOCOM 2017-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Piscataway:IEEE Press,2017:559-564.

Traffic classification algorithm of Internet of things devices based on random forest

LI Ruiguang^{1,2}, DUAN Pengyu¹, SHEN Meng¹, ZHU Liehuang^{1,*}

(1. School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China;

2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: The traffic classification of Internet of things (IoT) devices is very important to the management of cyberspace assets. The classification technology based on statistical identification is a hot spot in current academic research. The previous algorithms were mainly based on the flow information to set up the feature vectors, but less on the packet information. In this paper, we improve the traffic classification algorithm of IoT devices based on random forest. We set up the feature vectors with both the flow information and the flow's packet information. The experimental results show that, compared with previous algorithms, the classification accuracy of the proposed algorithm increases from 56% to 82%, the recall rate improves from 47% to 67%, the F_1 score increases from 0.43 to 0.74, and the confusion matrix correlation is also significantly improved. As a result, the proposed algorithm has a better classification effect than previous ones.

Keywords: Internet of things (IoT); traffic classification algorithm; random forest; feature vector; flow information; packet information

http://bhxb.buaa.edu.cn jbbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0401

大规模物联网恶意样本分析与分类方法

何清林^{1,2}, 王丽宏^{2,*}, 罗冰¹, 杨黎斌³

(1. 国家计算机网络应急技术处理协调中心, 北京 100029; 2. 北京航空航天大学 计算机学院, 北京 100083;
3. 西北工业大学 网络空间安全学院, 西安 710072)

摘 要: 物联网 (IoT) 恶意样本发展迅猛, 在网络中大量攻击各类物联网设备, 但由于开源问题导致其家族特征并不明显, 需要一种更细粒度的样本分类方法, 以解决高级威胁样本发现和攻击组织追踪等问题。针对该问题, 对 2019 年 5 月至 2020 年 5 月捕获到的 157 911 个物联网恶意样本进行了大规模分析, 并标注了一套包含 9 个家族分支共计 12 278 个样本的数据集。提出了物联网恶意样本的分类方法, 通过静态逆向分析提取 FCG 图和文本等复杂结构特征, 利用图表示学习和文本表示学习的特征, 在标注的数据集上取得了平均召回率 88.1% 的分类效果。所提方法在实际工作应用中效果优异。

关 键 词: 物联网 (IoT); 恶意样本; 分类; 图学习; 文本学习

中图分类号: TP393.4; TP312

文献标志码: A **文章编号:** 1001-5965 (2022) 02-0240-09

据世界经济论坛 (WEF) 发布的《2020 年全球风险报告》^[1], 网络攻击将成为继气候灾难、极端天气、核危机等之后人类面临的第七大威胁, 另据统计机构 Statista 的预测报告^[2], 物联网设备将成为网络中的主流设备, 2025 年将超过 750 亿台, 各类物联网设备已经成为网络攻击的重要目标。2016 年 9 月, 第一个出现的大型物联网僵尸网络 Mirai, 其恶意样本的功能和结构相对简单, 之后各种各样的物联网恶意样本不断涌现, 它们的功能越来越多, 攻击手法越来越复杂, 感染各类物联网设备的能力也越来越强, 已经成为网络空间中的重要威胁。物联网恶意样本含有攻击模块等漏洞利用相关的代码, 以及 C&C 服务器地址、代码编写风格等体现作者属性的信息, 对样本进行分析、分类, 可以获得关键线索, 支撑未知威胁发现、攻击组织溯源等工作。

传统恶意样本一般使用家族标签对其进行分

类, 但是对物联网恶意样本来说, 家族标签并不适用。首先, 物联网恶意程序现有的家族标签信息并不准确, 目前已知的 2 个最大物联网恶意样本家族是 Mirai 和 Gafgyt, 但是这两者都开放了源代码, 后续的大部分物联网恶意样本都混合使用了其源代码, 无法实际区分两者。其次, 物联网恶意样本的家族标签信息不够, 无法精确刻画恶意样本的类别, 以 Mirai 和 Gafgyt 为基础延伸出多个分支的恶意样本, 仅使用家族标签无法对其进行精确刻画和分类。例如, 广泛传播的 Mozi 系列样本, 就是一个融合了 P2P、Mirai、Gafgyt 等多种技术的系列样本, 但是并没有明显的家族标签与之对应。因此, 需要一种更精确的物联网恶意样本分类方法, 对样本的各个家族分支进行追踪和分析。

为了解决该问题, 本文开发了一套物联网恶意样本捕获与分析系统, 在 2019 年 5 月至 2020 年

收稿日期: 2020-08-09; 录用日期: 2020-09-05; 网络出版时间: 2021-03-18 08:40
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210317.1134.002.html
基金项目: 国家重点研发计划 (2017YFC1201204)
* 通信作者. E-mail: wlh@isc.org.cn

引用格式: 何清林, 王丽宏, 罗冰, 等. 大规模物联网恶意样本分析与分类方法[J]. 北京航空航天大学学报, 2022, 48 (2): 240-248. HE Q L, WANG L H, LUO B, et al. Large-scale IoT malware analysis and classification method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48 (2): 240-248 (in Chinese).

5 月在某运营商的网络出入口节点上共捕获到了 157 911 个在野传播的物联网恶意样本。通过对样本进行静态逆向和多种维度的统计分析,结合分析结果和关联的威胁情报数据,标注了一套物联网恶意样本数据集,并提出了一种新的物联网恶意样本分类方法。该分类方法能够实现跨平台架构和高效的物联网恶意样本分类任务,并且在标记数据集上取得了优异的分类效果,平均召回率达 88.1%。

本文的主要贡献如下:

- 1) 对在野传播的 157 911 个物联网恶意程序进行了分类和梳理,对其 CPU 架构分布、加壳方式、样本传播方式等进行了分析,给出了物联网恶意样本的最新流行趋势。
- 2) 结合多维度情报数据,对部分物联网恶意样本进行了标注,形成了一套物联网恶意样本数据集。该数据集包括 9 个家族分支,共计 12 278 个样本,覆盖 X86、MIPS、ARM 等 3 种 CPU 架构。
- 3) 提出了一种新的物联网恶意样本的分类方法。该方法通过静态逆向分析提取属性函数调用图 (attributed functions call graph, AF CG) 和字符串文本等样本信息,进一步使用图表示学习和文本表示学习的无监督学习算法,提取高维的向量特征,实现对跨 CPU 架构物联网恶意样本的统一表示。
- 4) 对学习到的高维向量特征进行了分类实验验证,使用有监督学习分类支持向量机 (SVM) 算法进行分类,在标注数据集上达到平均 88.1% 的召回率。实验表明,从样本提取出能表征样本特征的信息越多,越能够明显提高分类的准确率。

1 相关研究

2016 年,Mirai 事件的爆发让物联网恶意样本进入人们的视野,Antonakakis^[3]、de Donno^[4]等对 Mirai 事件过程及样本的机理进行了详细描述。Mirai 初期主要使用内置密码的密码表,对大量具有弱口令的物联网设备进行了密码爆破攻击,在高峰时期控制了多达 60 万台的物联网设备。

物联网恶意样本主要是 Linux 平台下的 ELF 文件,Cozzi 等^[5]对 10 548 个 ELF 恶意文件进行了系统性分析,结合使用静态分析和动态分析的方法,分析了恶意样本的 ELF 文件头、持久化、加壳、信息收集、隐藏手法等 9 类行为特征,并指出

同一家族的恶意代码会具有明显不同的行为特征。

Herwig 等^[6]分析了一个新的物联网恶意程序家族 Hajime,使用人工样本逆向分析和流量分析的方法获知 Hajime 的样本行为,并估算僵尸网络的规模,该僵尸网络是披露的第一个大规模 P2P 物联网僵尸网络。文献[7]分析了另外一个 2020 年新出现的大规模 P2P 僵尸网络 Mozi,病毒检测引擎 VirusTotal^[8]目前把 Mozi 样本识别为 Mirai 和 Gafgyt,还没有新的家族标签。

文献[9-10]提出了基于卷积神经网络 (CNN) 深度学习的恶意程序判定和分类方法,其中文献[9]将物联网样本文件直接切割和转换,变成灰度图片后用 CNN 的模型进行恶意性判定,文献[10]根据样本文件构造熵灰度图后使用 CNN 深度学习模型进行分类。这类方法都没有对样本文件本身进行深入的分析,没有还原样本的结构特征和语义特征,可解释性不足。

物联网恶意样本是编译链接好的二进制文件,与传统的 Windows PE 样本及 Android APK 样本相比,有很多相似之处,其结构特征一般包括逆向还原的函数调用图 (functions call graph, FCG)、控制流图 (control flow graph, CFG) 等,语义特征一般包括函数的反汇编代码片段等。恶意样本文件特征提取是一项繁杂的工作,涉及复杂的静态逆向分析,IDA Pro 是最常用的静态分析工具,文献[11]对如何用 IDA Pro 分析物联网恶意样本给出了总结。其他常用的逆向分析工具还有 Angr^[12]、Radare2^[13]等。

文献[14-15]提出使用 FCG 作为恶意样本相似性比较的主要参数,其中文献[14]使用优化的图编辑距离算法计算 2 个 FCG 图之间的距离,文献[15]则使用更多维度的函数属性值来作为 FCG 的节点属性,并采用学习优化的方法来计算 FCG 的距离矩阵。

可以看出,提取样本文件本身的复杂结构特征和语义特征,并对其进行表征和学习,是恶意样本研究的主流方向。物联网恶意样本出现时间不长,与传统 Window 样本和 Android 样本相比又有很多新的特性,如支持的 CPU 架构平台多样等,目前专门针对物联网恶意样本的分类研究还较少。本文借鉴传统的样本特征提取方法,同时考虑到物联网样本多平台问题,提出一种结合图和文本特征,以及结合无监督学习和有监督学习的样本分类方法。

2 物联网恶意样本分析

2019 年 5 月至 2020 年 5 月,在某运营商网络中一共捕获到 157 911 个传播的物联网恶意样本。通过对这些恶意样本进行大规模静态逆向分析,包括其支持的 CPU 架构、加壳方式、漏洞利用传播方式等,形成了一套物联网恶意程序标注数据集。分析流程如图 1 所示。

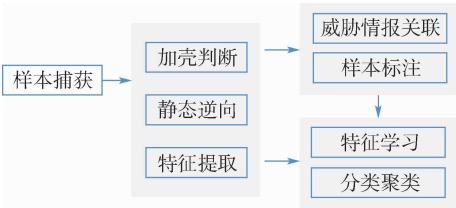


图 1 样本分析处理示意图

Fig. 1 Schematic diagram of sample analysis and processing

2.1 样本 CPU 架构分布

虽然物联网设备的硬件形态及 CPU 架构多种多样,但是恶意程序的主要攻击目标是操作系统基于 Linux 内核的设备,因此物联网恶意样本的形态主要为 Linux 下的 ELF 可执行文件。黑客在制造恶意程序时,一般都会交叉编译成多个版本的 ELF 文件同时发布,以支持不同的 CPU 架构。根据所有样本的统计,其支持架构分布如表 1 所示。

统计数据显示,物联网恶意样本攻击最多的为 ARM 和 MIPS 设备。可能因为在互联网中暴露且经常会出现漏洞的物联网设备主要是路由器和摄像头两大类,其 CPU 架构主要也是这 2 种,容易被黑客所青睐。另外统计数据还显示,大部分支持 Sparc、PowerPC 等架构的样本,同时也有支持 ARM、MIPS 和 X86 等架构的样本存在。这说明很多的物联网恶意程序会同时发布 7~8 个版本。发现某个 CPU 架构的恶意样本后,在恶意

表 1 157 911 个恶意样本 CPU 架构分布

Table 1 Distribution of CPU framework of 157 911 malwares		
CPU 架构	样本个数	占比/%
ARM	45 663	28.9
MIPS	28 569	18.1
X86	16 956	10.7
MC68000	11 576	7.3
PowerPC	10 202	6.5
SuperH SH	9 793	6.2
Sparc	7 703	4.9
X86-64	5 210	3.3
其他	22 239	14.1

样本库中快速找到跨平台的其他同源恶意程序,也是大规模样本分析的一个基本目标。

2.2 样本加壳

通过对恶意样本加壳进而隐藏相关信息,防止被安全人员逆向分析,是黑客常用的一种对抗手段。大规模样本分析特别是静态分析,需先解决样本加壳问题,然而对恶意程序进行脱壳从来就不是一项容易的工作。Windows 平台下的 PE 样本和 Android 平台下的 APK 样本,由于发展时间较长,存在各种各样的加壳方式,脱壳对抗难度非常复杂。物联网平台下的 ELF 样本,由于大规模出现的时间相对较晚,情况稍有不同,目前主要采用相对单一的 UPX 方式进行加壳。

UPX 本身是一种开源的文件压缩方式,恶意程序编写者们会通过改写 UPX 文件头的各种参数来防止解压。文献[5]中对 2017—2018 年捕获的 10 548 个 ELF 样本进行统计,发现只有 3.8% 的恶意程序使用了 UPX 加壳。本文对 2019 年 5 月至 2020 年 5 月捕获的 157 911 个恶意样本进行分析发现,其中有超过 37% 的恶意样本采用了 UPX 加壳,详情如表 2 所示。

因此从长期来看,黑客会采用更多样、更复杂的方式对 ELF 文件加壳,这将是安全人员长期面临的一项对抗与挑战。

表 2 物联网恶意样本加壳方式

Table 2 Packers of IoT malware		
壳类型	样本个数	占比/%
UPX 变种	41 026	26.0
标准 UPX3.94	15 373	9.7
标准 UPX3.95	2 148	1.4
其他 UPX 标准版本	623	0.40

2.3 传播方式

早期的物联网恶意样本,如 2016 年在美国爆发的 Mirai,主要采用 SSH 密码爆破的方式进行植入和传播。随着黑客技术的演进和攻防对抗的升级,越来越多的物联网恶意程序开始利用设备的漏洞进行植入和传播。2019 年 5 月披露的 Echo-bot 系列样本^[16]使用了多达 70 个已知的漏洞利用进行传播。2020 年 3 月,笔者团队发现有 Fbot 的恶意样本开始利用 0-day 漏洞进行传播。

针对所有捕获的样本进行漏洞利用分析,在其中的 68 648 个样本中发现有已知漏洞利用行为,共计发现超过 200 种物联网漏洞利用行为,其中最常用的物联网漏洞有 CVE-2017-17215 等。统计前 10 个利用最多的漏洞,详情如表 3 所示。

表 3 样本 10 大漏洞利用统计

Table 3 Statistics of top 10 vulnerability exploited in malware

漏洞	影响设备	发现样本个数
CVE-2017-17215	华为 HG532 家用路由器	18 025
CVE-2014-8361	使用 Realtek SDK 摄像头	7 108
CVE-2017-6884	Zyxel 家用路由器	6 945
Redis 命令执行	安装 Redis 服务的设备	5 433
CVE-2018-10561	GPON 路由器	2 369
JAWS 命令执行	VPower DVR 等	1 804
Vacron 命令执行	Vacron NVR 设备	1 534
Linksys 命令执行	Linksys 系列路由器	1 281
Dlink 命令执行	Dlink 系列路由器	1 250
Netgear 命令执行	Netgear 系列路由器	837

另外,笔者团队发现物联网恶意样本在网络中传播时会使用特定的文件名,如 Hajime 家族^[6]主要采用“.i”的文件名进行传播,最新的 Mozi 系列样本^[7]使用了“Mozi. a”和“Mozi. m”的文件名。图 2 显示了 2 个典型 Mozi 样本的逻辑结构,分别是“Mozi. a”和“Mozi. m”,具有不同的 CPU 架构,但是具有相似的逻辑结构。

样本传播方式包括传播文件名和漏洞利用手法,是标注数据集时参考的一项主要参数。

2.4 标注数据集

在恶意样本识别和分类研究中,目前使用最多的数据集是 2015 微软恶意样本分类挑战赛数据集^[17],其中的恶意程序主要是 Windows 平台下

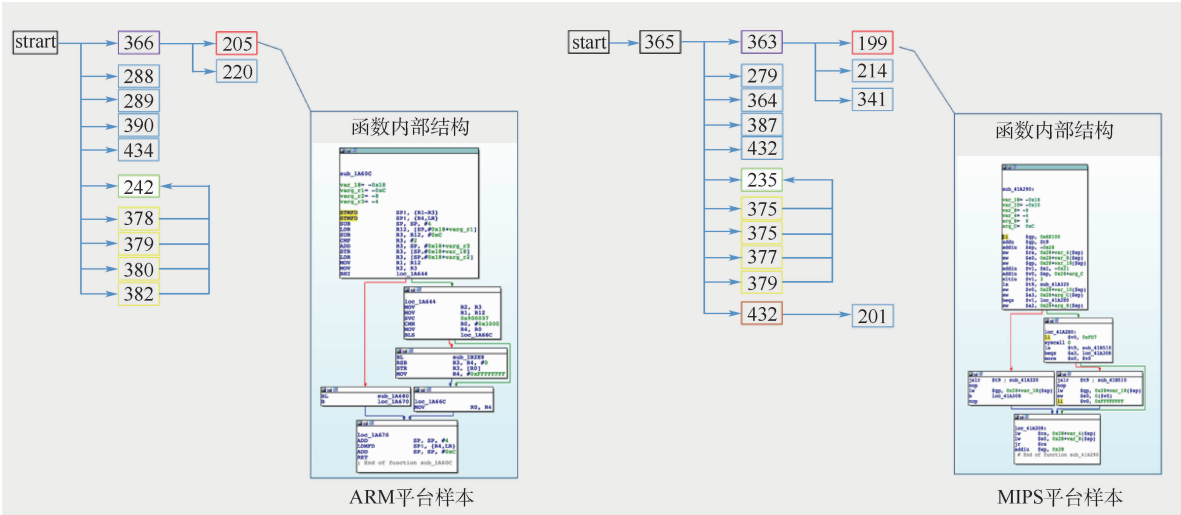


图 2 两个 Mozi 样本的 FCG 示意图

Fig. 2 Schematic diagram of FCG of two Mozi samples

的 PE 文件,包含 9 类共 21 704 个恶意程序。而针对物联网恶意样本的研究,目前并没有一个公认的数据集。

根据捕获到的恶意样本文件,以及相关的传播地址、C2 地址等威胁情报信息,同时参考病毒检测引擎平台 VirusTotal^[8]的样本检测结果和部分人工分析结果,标注了一套物联网恶意样本数据集,作为后续进行大规模样本分类研究的基础。如此规模的数据集标注,无法依赖人工分析手动完成,须通过自动识别和关联,因此数据集的标注要求至少满足以下 2 个条件:①样本能够通过程序自动脱壳并成功提取到静态编译信息,以满足后续的特征处理要求;②样本有相关的关联威胁情报信息,如样本使用的传播文件名和传播地址、样本在 VirusTotal^[8]上有检索结果等信息,用来确定该样本是哪个具体的家族或者分支。但是在捕获的所有样本中,其中有大部分样本或是无法进行自动脱壳,或是无法关联到相关信息,因此无法

对其进行标注并纳入到标注数据集中。最终从 157 911 个物联网恶意样本中成功标注了 12 278 个样本,作为标注数据集进行分类实验。包括 9 个类别,覆盖 ARM、MIPS、X86 三种 CPU 架构,如表 4 所示。数据集中的样本都是已脱壳并能通过反编译工具提取出静态特征信息。

数据集中的类别与恶意样本的家族并不是一一

表 4 物联网恶意程序标注数据集

Table 4 Labeled IoT malware dataset

类别	样本个数	覆盖 CPU 类型	类别说明
1	241	ARM, MIPS, X86	Echobot 系列样本
2	123	ARM, MIPS	Mozi 系列样本
3	2 676	ARM, MIPS, X86	UnHAnaA 系列样本
4	598	ARM, MIPS, X86	JoSho 系列样本
5	4 501	ARM, MIPS, X86	Loligang 系列样本
6	2 185	ARM, MIPS, X86	Yakuza 系列样本
7	807	ARM, MIPS, X86	Sora 系列样本
8	234	ARM, MIPS, X86	Fbot 系列样本
9	913	ARM, MIPS, X86	Owari 系列样本

一对应关系,而是某一家族样本的某类具体分支。即使同一个样本,在 VirusTotal 上不同引擎也会给出不同的家族标签,而目前物联网恶意代码的家族判定并没有统一的标准。本文目标不是研究样本的恶意性判定或者家族判定,而是研究如何对已发现的大量 IoT 恶意样本进行更详细的分类,找出具有同类行为的分支,并对其进行更深入的追踪,从而溯源攻击组织、发现具有高级威胁的未知样本等。

3 样本特征提取

恶意样本分析,一般来说有静态分析和动态分析 2 种方法。静态分析主要通过反汇编反编译,提取二进制文件的语义和语法信息;动态分析则是通过在虚拟环境下运行可执行程序,获得各

类运行状态行为信息。目前,物联网恶意程序支持的设备种类众多,缺乏成熟的动态执行环境。本文主要采用静态分析的方法提取函数调用关系图、函数反汇编代码、字符文本等特征,主要分析流程如图 3 所示。

3.1 函数调用图

FCG 是能够从整体上直接反应恶意样本逻辑结构的一种表示。它是一类典型的图结构 $G(\text{node}, \text{edge})$, node 表示恶意样本中的函数, edge 表示函数之间的调用关系, edge 是有向边。图 2 显示了 2 个同分支的样本,分属不同的 CPU 架构,但具有相似的 FCG 结构和函数结构。图中:节点编号代表了函数的排列序号,样本总计有 400 多个函数节点,为方便表示只标记了从起点函数开始的前 3 层调用节点。

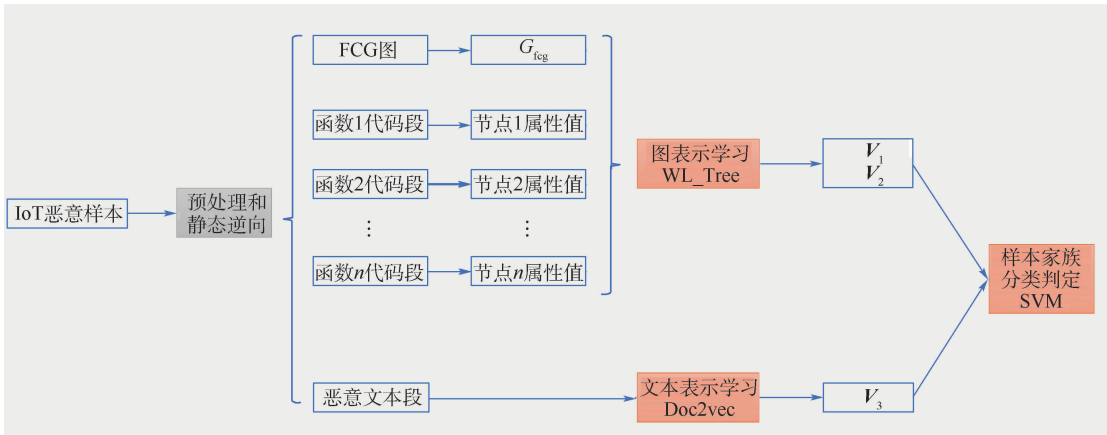


图 3 样本特征提取与分类方法示意图

Fig. 3 Schematic diagram of sample feature extraction and classification method

能否成功从样本中还原出 FCG 图,取决于 2 个因素:①样本没有加壳或者能够成功解壳;②样本没有使用外部动态链接库。本文数据集都进行了脱壳处理,而且都仅使用静态链接的样本。另外,根据对物联网样本库的统计,目前仅有 3.2% 的样本使用了动态链接库。

对恶意样本直接进行静态逆向得到的 FCG 图,通常不是一个连通图。例如,很多物联网的恶意程序都有一个初始化函数 init_func,其表示的节点在图中是个独立的节点。根据对样本库的统计数据显示,最大连通子图中的节点个数占整个图的节点个数比率的分布区间为 93.2% ~ 99.1%,平均值为 96.7%。选取 $G(\text{node}, \text{edge})$ 的最大连通子图作为恶意程序的函数调用图近似表示 G_{fcg} ,可以认为 G_{fcg} 就是恶意程序的逻辑结构表示。

本文针对数据集集中的 12 278 个不同的恶意程序文件进行分析提取,得到 1 323 个不同的 G_{fcg} 文件。这表明很多恶意样本具有同样的 G_{fcg} ,样

本的逻辑结构是完全一致的,恶意样本的编写者们只是进行了一点简单修改,如修改 C2 地址或对某个函数的功能进行升级等,再重新编译作为一个新的恶意程序发布。以数据集集中的第 4 类 Josho 样本为例,有 2 个 MD5 值不同的样本,其样本 MD5 值分别为:“aeb9d28e524694e5ef5eb6275ef53b05”和“90e2ab9037c9b18cb2ea83c5f3862ee6”(VirusTotal^[8]上均可以检索到样本相关信息),文件大小分别为 56 584 Byte 和 58 180 Byte。通过静态分析提取得到的这 2 个样本的函数调用关系 G_{fcg} 完全一样,而这样的情况在同一类样本中大量出现。由此可以看出, G_{fcg} 图是一种能够快速识别和分类相似恶意样本的方法。

进一步,如果加上节点属性,也就是函数的属性,就可以得到恶意程序的逻辑结构及语义表示,形成属性函数调用图,用 G_{afcg} 表示。

3.2 函数属性

恶意样本是一类可执行的二进制文件,而函

数是二进制文件的基本逻辑单元。针对二进制文件中的函数属性提取和学习是逆向静态分析的一个重要研究方向,尤其是漏洞挖掘样本关联分析等。很多安全分析人员直接根据函数的 hash 值进行分析^[18];也有使用函数控制流图信息来进行大规模二进制分析^[19],如漏洞挖掘固件组件分析等;文献[20]使用 BERT 深度学习框架学习函数的特征。

与以上方法不同,本文使用一种相对高效且跨平台的方式来学习函数的向量化表示。先对恶意程序进行静态逆向得到函数的一段反汇编码。在反汇编代码中,一般包含 API 调用(系统或外部调用)和本地汇编操作。一个典型的本地汇编操作包含操作指令和地址或变量。其中地址或变量是动态执行相关的,通过静态逆向分析得到的值参考意义有限,因此重点关注操作指令。把函数转换成一段包含操作指令和 API 调用的序列。

不同 CPU 架构的指令集规模不一样,统计 157 911 个物联网恶意样本用到的指令,按照 CPU 架构类别统计,如表 5 所示。更进一步,为了解决函数属性跨平台统一表示的问题,参考文献[21-22]的方法,将操作指令统一为中间语言表示。定义 20 个不同的操作类型,根据指令类别和在样本中出现的频率,将不同 CPU 架构的指令映射到这 20 类中,将 API 调用映射为另外 1 类。目前,仅对物联网恶意样本中最常见的 3 种 CPU 架构,也是样本集中涉及到的 3 种 CPU 架构,即 ARM、MIPS、X86 做了映射和统一表示。

提取的函数属性值是一段有序的数值序列 A,序列中的数值为 1~21 的数。通过对序列 A 学习的向量作为函数的属性值。进一步,将函数与 G_{afcg} 中的节点对应起来,并将函数属性值作为节点的属性值,得到样本的 G_{afcg} 表示。从样本中一共提取出了 1 323 个不同的 G_{afcg} 文件。

表 5 恶意样本中常用的 3 种 CPU 架构指令集数量
Table 5 Number of instruction set commonly used in malware with three CPU frameworks

CPU 类型	样本中使用的指令集数量
ARM	179
MIPS	198
X86	195

3.3 字符串信息

物联网恶意程序在编译连接生成 ELF 格式的可执行文件时,会将很多字符串一起打包到可执行文件中。这些字符串一般包括本地的操作命令、密码表、远程通信命令等信息。通过反汇编提

取恶意程序可执行文件的字符信息,从而来判断恶意程序的行为,是逆向分析时的一个基本手段,如提取恶意程序外联的 C2 地址、恶意程序传播时利用的漏洞、使用的爆破密码表等。字符串包含了丰富的信息,对其内容进行表示学习和分类,可以更好地对恶意程序进行识别和分类。

从恶意程序提取出来的字符串是按行排列的,每一行代表一类操作或者一类命令,行与行之间是上下文无关的。对样本提取原始的字符串,再进一步重排序处理,得到 8 590 个不同的字符串文本文件 malware_string。

样本集中的样本个数是 12 278 个,提取出来的不同的 G_{afcg} 文件是 1 323 个,不同的字符串文本文件是 8 590 个。字符串文本文件和 G_{afcg} 文件的数目并不一致,多个样本可能对应同一个 G_{afcg} 文件,却对应不同的字符串文本文件。因为同一个样本进行修改再编译发布时可能不改变函数逻辑结构 FCG,却会修改字符串文本文件,所以样本分类时,将主要参考 FCG 图和 AFCG 图等逻辑结构信息,字符串文本特征作为辅助。进一步对字符串文本进行映射和处理后,从样本集中一共提取到 1 323 个不同的字符串文本文件。

4 特征学习及分类实验

从物联网恶意样本中提取出来的特征是包括图和文本等信息的复杂结构特征,无法直接使用这些特征来进行恶意样本的分类任务。因此,使用图表示学习和文本表示学习的方法,将恶意样本的特征转换成统一的向量表示。

针对实际场景设计了 9 组分类方法实验,验证从恶意样本中提取和学习到的特征的有效性。

4.1 图表示学习

文献[14-15]用恶意样本的 FCG 图进行分类判定,两两计算 2 个图之间的编辑距离,再以此为基础构建和优化相似性模型。本文做法与之前的方法不同,使用图表示学习的方法,将 FCG 图通过表示学习直接映射到一个向量空间中。

图表示学习,目前主要有基于核函数的方法和图神经网络方法两大类。图神经网络方法更适合学习节点的特征用于节点分类任务,而针对整个图的分类任务,则使用图核函数的方法更多。目前,常见的图核函数有基于随机游走的核函数、基于路径的核函数、基于子树结构的核函数等。

对图的整体结构捕获和表达能力最好的是基于子树结构的 WL_Subtree 核函数^[23],其是一种快速计算的子树核,由多步迭代计算完成。

图表示学习对象为:无节点属性的 FCG 图和包含节点属性的 AFCG 图。从样本数据集中一共提取出 1 323 个不同的图,图的节点规模分布区间为 212 ~ 3 226,均值为 372。对 FCG 图 G_{fcg} 和 AFCG 图 G_{afcg} 分别进行图表示学习,得到 2 组特征向量 V_1 和特征向量 V_2 ,如式(1)、式(2)所示,其中 $k_{WL_SubTree}$ 表示 WL_Subtree 核函数:

$$k_{WL_SubTree}(G_{fcg}) = V_1 \tag{1}$$

$$k_{WL_SubTree}(G_{afcg}) = V_2 \tag{2}$$

4.2 文本表示学习

目前,文本学习方法主要有轻量级的 TF-IDF、Word2vec^[24] 等简单模型,也包括像 BERT^[20] 深度学习框架等较为复杂的模型。TF-IDF 实现相对简单,较为容易理解,不足在于没有考虑特征在样本间分布情况,难以适用于恶意样本分类场景。BERT 深度学习框架能够梳理出连续文本的内在联系和语言结构,但对数据规模及模型大小要求非常苛刻,一般适用于大规模海量的数据场景中。

学习对象为恶意样本提取出来的字符串形式的文本标记为 malware_string,其大小分布一般在 5 KByte 至 50 KByte 之间,样本中需要用来进行下游分类任务的文本 malware_string 的数量为 1 323 个。综合性能和效率考虑,同时能够更好地利用隐藏在样本序列中的语义和语法信息,引入 Doc2vec^[25] 方法进行向量特征提取。Doc2vec 方法是一种无监督算法,可从变长的恶意样本字符串文本中学习得到固定长度的特征表示,既可克服词袋模型中没有语义的缺点,又能接受不同长度的恶意样本做训练,并表征为向量。利用 Doc2vec 方法进行向量特征学习过程,主要包括去噪、训练及推断等 3 个步骤,如图 4 所示。

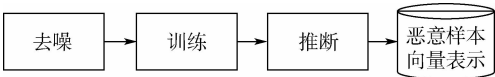


图 4 基于文本表示学习的向量特征学习

Fig. 4 Vector characteristic learning based on text representation learning

在 Doc2vec 方法中,首先在去噪过程中通过预处理剔除了一些恶意文本中噪声数据,将恶意文本库的格式修剪统一;其次在训练过程中,主要期望在已知的恶意文本训练数据中得到恶意字符串的词向量、softmax 的参数、恶意文本段落向量或句向量以进行后续推断操作;最后在推断过程中,主要利用已有信息,推断后续新的段落向量表达形式,得到恶意文本的向量表达形式。最终通过 Doc2vec 方法学习到的恶意文本向量表达记为特征向量 V_3 ,如下:

$$\text{Doc2vec}(\text{malware_string}) \rightarrow V_3 \tag{3}$$

4.3 二分类学习

设计的分类实验是二分类学习,对应的实际场景是:如果已知某一物联网恶意样本家族分支的系列样本后,能否通过学习的方法,快速地从样本库中或者新发现的样本找出同类的样本。将样本集中的每一类样本分别标记为正例样本,样本集中的其他样本标记为负例样本,做二分类学习和判定。分类算法使用 SVM。

SVM 是基于超平面分类线性可分的假设,通过学习和优化得到一个超平面分类器。假设有 N 个训练集 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, $x_i \in \mathbf{R}^d, y_i \in \{-1, 1\}$ 。假设空间为超平面 (a, b) , a 为权重向量, b 为偏差,对一个新输入的数据 x 可以用如下分类:

$$f(x) = \text{sign}(ax + b) = \text{sign}\left(\sum_{i=1}^N a_i y_i (x_i \cdot x) + b_i\right) \tag{4}$$

该分类问题是个不平衡分类问题,正例样本的个数远小于负例样本的个数,实际场景中也是如此。因此,更关注正例样本是否被正确标记,把召回率 R 作为主要参考的评价指标,同时参考 F_1 值,指标如式(5)和式(6)所示。其中 TP 表示被正确标记正例样本,FP 表示被错误标记的正例样本,FN 表示被错误标记的负例样本。

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{6}$$

选 3 种不同的样本特征作为分类学习的算法输入,分别使用从样本 FCG 图学习到的向量 V_1 (1)、从样本 AFCG 图向量学习到的向量 V_2 (2)、 V_2 和从样本的恶意文本学习到的向量 V_3 (3) 相结合。不同的特征选择,代表了不同的样本表征能力。实验中,将训练集和测试集划分为 0.8 和 0.2,最终选取多次实验的平均值作为最终指标。实验结果如表 6 所示。

表 6 不同特征和类别的分类召回率及 F_1 值

Table 6 Classification recall rate and F_1 value with different features and categories

类别	V_1		V_2		$V_2 + V_3$	
	$R/\%$	F_1	$R/\%$	F_1	$R/\%$	F_1
1	100	1.0	100	1.0	100	1.0
2	87	0.91	90	0.92	92	0.93
3	85	0.92	89	0.93	90	0.93
4	47	0.52	67	0.72	70	0.74
5	89	0.91	92	0.94	94	0.95
6	93	0.93	94	0.94	96	0.96
7	93	0.97	93	0.97	94	0.97
8	82	0.84	86	0.89	88	0.91
9	59	0.65	68	0.73	69	0.73
均值	82	0.85	87	0.89	88.1	0.92

4.4 实验结果分析

实验结果显示,样本分类平均召回率达到 88.1%,说明从样本中提取的特征是有效的,通过表示学习得到的向量,可以很好地从整体上表征恶意样本。实验结果还显示,使用的特征越多,分类效果越好,说明如果能够从样本中提取出更多能表达样本逻辑结构和语义信息的特征,更有助于提高分类的效果。

实验中的 9 组分类实验,分别对应数据集中 9 个不同的样本家族分支类别。其中除了第 2 组 Mozi 分支以外,其他 8 个分支均由 Mirai 家族演化而来。其中第 4 组和第 9 组的样本分类召回率较低,分别是 JoSho 家族分支的 74% 和 Owari 家族分支的 73%。对这 2 个家族分支的部分样本进行人工逆向分析,发现其样本中大量借鉴了 Mirai 的开源源码。例如,属于 JoSho 家族分支的 MD5 值为“aeb9d28e524694e5ef5eb6275ef53 b05”的样本,和属于 Owari 家族分支的 MD5 值为“46a4cea56dafbfecd0981a7d411508c9”的样本(VirusTotal^[8]上均可以检索到样本相关信息),通过对这 2 个样本进行人工逆向分析发现,其与 Mirai 源代码的差异性不到 10%。而这 2 个样本在这 2 类家族分支具有典型性,在该家族分支中分别存在多个与这 2 个样本相似的样本。这说明这 2 个家族分支在样本中自己进行编码的部分并不多,也就是说没有很明显的样本分支特征,从而导致分类效果不是很理想。

实验结果显示,其余 7 组家族分支的样本检出率均超过了 0.8,其中 1 组的检出率为 1.0,说明本文样本分类方法能够对大部分的样本家族分支进行正确地分类,实现对不同家族分支的样本进行追踪分析的目标。

5 结 论

目前,针对物联网恶意样本的大规模分类研究属于刚刚起步阶段,本文根据实际工作中捕获到的物联网恶意样本,进行了大规模分析和分类的尝试:

- 1) 对所有样本进行了分析和统计,并结合外部威胁情报特征和人工分析,标注了一套物联网恶意样本数据集。
- 2) 通过静态逆向分析提取特征和学习特征,在标注的样本数据集上进行分类实现,效果表现优异,达到了平均 88.1% 的召回率。
- 3) 该样本分类方法在实际工作中取得了较好的效果,能够持续地自动标注新的样本。

在实际工作中,笔者一直在持续跟踪 Mozi 系列的样本^[7],即标注数据集中的第 2 类,通过数据集训练生成的分类器,从原来的样本库中识别出了数十个之前未曾被标记的 Mozi 系列样本,并通过后续进一步的人工分析一一确认,通过分类器自动标记的 Mozi 样本是有效的。进一步,该方法已经用在后续的日常样本分析中,将每天新发现的样本进行自动分类和标注。

恶意样本分析和分类与其他领域(如图像识别分类等)不同,其研究对象恶意样本是由黑客制造出来,对象的性质会由于黑客持续对抗而变化。对样本的特征提取会由于加壳等技术的升级而变得日益困难,特征的表达能力可能也会由于使用代码混淆等高级对抗手法而受到影响。笔者将深入挖掘样本多个维度的特征,并结合安全专家知识,对物联网恶意样本分类问题进行持续研究。

参考文献 (References)

[1] World Economic Forum. The global risks report 2020 [EB/OL]. (2020-01-15) [2020-07-03]. <https://www.weforum.org/reports/the-global-risks-report-2020>.

[2] Gartner Inc. Gartner identifies top 10 strategic IoT technologies and trends [EB/OL]. (2018-11-07) [2020-07-03]. <https://www.gartner.com/en/newsroom/press-releases/2018-11-07-gartner-identifies-top-10-strategic-iot-technologies-and-trends>.

[3] ANTONAKAKIS M, APRIL T, BAILEY M, et al. Understanding the Mirai botnet [C] // USENIX Security Symposium, 2017: 1093-1110.

[4] DE DONNO M, DRAGONI N, GIARETTA A, et al. DDoS-capable IoT malwares: Comparative analysis and Mirai investigation [J]. Security and Communication Networks, 2018, 2018: 7178164.

[5] COZZI E, GRAZIANO M, FRATANTONIO Y, et al. Understanding Linux malware [C] // IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2018: 161-175.

[6] HERWIG S, HARVEY K, HUGHEY G, et al. Measurement and analysis of Hajime a peer-to-peer IoT botnet [C] // Network and Distributed Systems Security Symposium, 2019: 1-15.

[7] 国家互联网应急中心. Mozi 样本分析报告 [EB/OL]. (2020-02-28) [2020-07-03]. <https://www.ics-cert.org.cn/portal/page/112/f6aa66554f9a4669904d6b138cfe1ac.html>. CNCERT. Dive into Mozi malware [EB/OL]. (2020-02-28) [2020-07-03]. <https://www.ics-cert.org.cn/portal/page/112/f6aa66554f9a4669904d6b138cfe1ac.html> (in Chinese).

[8] Google LLC. VirusTotal [EB/OL]. [2020-07-03]. <http://virustotal.com>.

[9] SU J W, VARGAS D V, PRASAD S, et al. Lightweight classification of IoT malware based on image recognition [C] // IEEE 42nd Annual Computer Software and Application Conference. Piscataway: IEEE Press, 2018: 664-669.

[10] GIBERT D, MATEU C, PLANES J, et al. Classification of mal-

ware by using structural entropy on convolutional neural networks[C]//30th AAAI Conference on Innovative Applications of Artificial Intelligence,2018;1-6.

[11] SRI SHAILA G,DARKI A,FALOUTSOS M,et al. IDAPro for IoT malware analysis? [C]//Proceedings of the 12th USENIX Conference on Cyber Security Experimentation and Test, 2019;15.

[12] WANG F,SHOSHITAISHVILI Y. Angr-The next generation of binary analysis[C]//2017 IEEE Cybersecurity Development. Piscataway:IEEE Press,2017;8-9.

[13] Radare2[EB/OL]. [2020-07-03]. https://github.com/radareorg/radare2.

[14] HU X,CHIUEH T,SHIN K G. Large-scale malware indexing using function-call graphs[C]//ACM Conference on Computer and Communications Security. New York:ACM,2009;611-620.

[15] KONG D,YAN G H. Discriminant malware distance learning on structural information for automated malware classification[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM, 2013;1357-1365.

[16] CIMPANU C. New Echobot malware is a smorgasbord of vulnerabilities[EB/OL]. (2019-06-17) [2020-07-03]. https://www.zdnet.com/article/new-echobot-malware-is-a-smorgasbord-of-vulnerabilities.

[17] Microsoft malware classification challenge (BIG 2015) [EB/OL]. [2020-07-03]. https://www.kaggle.com/c/malware-classification.

[18] HARUYAMA T. fn_fuzzy: Fast multiple binary diffing triage with IDA [EB/OL]. (2019-05-09) [2020-07-03]. https://conference.hitb.org/hitbsecconf2019ams/sessions/fn_fuzzy-fast-multiple-binary-diffing-triage-with-ida/.

[19] XU X J,LIU C,FENG Q,et al. Neural network-based graph embedding for cross-platform binary code similarity detection [C]//ACM SIGSAC Conference on Computer and Communications Security. New York:ACM,2017;363-376.

[20] DEVLIN J,CHANG M,LEE K,et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT,2019;4171-4186.

[21] HEITMAN C,ARCE I. BARF: A multiplatform open source binary analysis and reverse engineering framework [C]//XX Congreso Argentino de Ciencias de la Computación,2014.

[22] ALAM S,HORSPOOL R N,TRAORÉ I. MAIL: Malware analysis intermediate language: A step towards automating and optimizing malware detection[C]//Proceedings of the 6th International Conference on Security of Information and Networks, 2013;233-240.

[23] SHERVASHIDZE N,SCHWEITZER P,VAN LEEUWEN E J,et al. Weisfeiler-Lehman graph kernels[J]. Journal of Machine Learning Research,2011,12;2539-2561.

[24] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of NIPS2013,2013;3111-3119.

[25] LE Q,MIKOLOV T. Distributed representations of sentences and documents[C]//Proceedings of ICML,2014;1188-1196.

Large-scale IoT malware analysis and classification method

HE Qinglin^{1,2}, WANG Lihong^{2,*}, LUO Bing¹, YANG Libin³

(1. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China;
2. School of Computer Science and Engineering, Beihang University, Beijing 100083, China;
3. School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Recently, Internet of things (IoT) malware emerges in large numbers and attacks IoT devices in cyberspace. However, the family characteristics of IoT malwares are not obvious due to the open-source problem, a more fine-grained malware classification method is needed to solve the problems of advanced threat malware discovery and attack organization tracking. To address this question, we took a large-scale analysis of 157 911 IoT malwares which have been found from May 2019 to May 2020, and labeled a dataset which includes 9 categories and 12 278 malwares. Then we proposed an IoT malware classification method whose main idea is extracting complex structure features including FCG graph and text by static reverse analysis. The learning features using graph representation learning and text representation learning were used, and the experiments on the labeled dataset show that the average recall rate is 88.1%. Our method has been taken into practice and works well.

Keywords: Internet of things (IoT); malware; classification; graph learning; text learning

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0403

面向异构大数据环境的数据脱敏模型

佟玲玲, 李鹏霄*, 段东圣, 任博雅, 李扬曦

(国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 不同场景下数据类型和脱敏需求的差异,使得传统的数据脱敏方法难以满足大数据背景下的用户隐私保护需求。如何实现异构大数据中敏感信息的精准定向、高效脱敏,从而更好地确保数据安全、可信和可用,是本领域的研究难点。提出了一种在异构大数据环境下,基于文本、图片、音频和数据库等异构数据的脱敏模型,并对 4 个关键模块进行了描述。通过脱敏数据预处理,实现不同应用场景下敏感数据的自动标注和分级设置。采用数据预脱敏处理方法,并从数据可用性、数据关联性、隐私保护度、时间和空间复杂度等 5 个维度进行脱敏效果评价,实现定制化脱敏策略。经过脱敏任务调度完成脱敏任务分配和执行,并支持用户对部分脱敏数据恢复。基于提出的异构大数据脱敏模型,对 2 种典型数据脱敏应用场景进行了验证分析,表明所提模型能够实现不同应用场景下异构敏感数据的高效脱敏。

关 键 词: 异构大数据; 敏感数据自动标注; 数据脱敏; 脱敏效果评价; 机器学习
中图分类号: TP391
文献标志码: A **文章编号:** 1001-5965(2022)02-0249-09

近年来,大数据分析技术已经广泛应用在国家治理、企业运行、个人日常生活等方方面面,数据成为重要的生产要素和最热门的基础资源,但同时海量的数据资源中往往蕴含有用户个人、企业,甚至国家重要行业的敏感、隐私信息,一旦遭到泄露或篡改,可能会给个人、企业甚至国家造成无法挽回的损失。如何在利用自身数据资源进行大数据研究分析的同时,避免敏感信息泄露的风险,成为了大数据分析应用领域众多研究者关注的热点问题。数据脱敏(data masking)又称为数据漂白、数据去隐私化或数据变形,是指在保留数据初始特征的前提下,通过脱敏规则对敏感数据进行数据的变形,避免未经授权的用户非法获取,实现敏感数据在分享和使用过程中的安全保护。数据脱敏可以在保存数据原始特征的同时改变其真实值,在保留数据有效性的同时保持数据的安全性,实现敏感隐私数据的可靠保护,避免敏

感数据泄露的风险^[1-6]。目前,相关研究者已经提出了多种方法用于解决数据脱敏相关问题,但主要集中在文本或数据库类型的数据,如 k -匿名(k -anonymous)^[1]、 l -多样性(l -diversity)^[7]、 t -保密(t -closeness)^[8]等,并对传统脱敏方法进行了改进,如 Sarada 等^[9]提出的基于最小最大归一化算法、范围映射脱敏算法,Gujjary 和 Saxena^[10]提出的基于神经网络的自适应脱敏算法,Zhou 和 Louis^[11]提出的基于空间平滑的矩阵屏蔽算法,吴克河等^[12]提出的基于敏感信息度量的 t -保密改良技术等。国内外研究主要关注具体脱敏方法并取得较多成果,但面向实际用的数据脱敏系统模型研究较少。Santos 等^[13]提出了一种针对数据库的脱敏模型,张琦颖^[14]和邵华西^[15]分别提出了基于 Spark 分布式计算框架下数据脱敏系统模型,但这类数据脱敏系统模型,在系统内置统一的脱敏规则实现数据库脱敏,或在脱

收稿日期: 2020-08-09; 录用日期: 2020-09-05; 网络出版时间: 2020-11-06 14:49
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201105.1659.005.html
基金项目: 国家自然科学基金(U1936110,U1836111)
* 通信作者. E-mail: lpx@cert.org.cn

引用格式: 佟玲玲, 李鹏霄, 段东圣, 等. 面向异构大数据环境的数据脱敏模型[J]. 北京航空航天大学学报, 2022, 48(2): 249-257. TONG L L, LI P X, DUAN D S, et al. Data masking model for heterogeneous big data environment[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 249-257 (in Chinese).

敏方法上的迁移性和普适性尚有欠缺。例如,某些脱敏方法重在数据保护,但会失去一些统计特征,适用于社交数据等数据冗余性高的场景;某些方法兼顾数据保护和可用性,但计算复杂度高,适用于医疗、金融等数据精度要求高的场景。

随着信息技术的不断发展,文本、图片、音频等异构数据量日益增大,大数据分析应用场景也日趋复杂,在实际生产环境或非可信环境下的数据脱敏需求也各不相同,如何在数据交换、共享及使用等过程中实现对敏感数据的精准定向、高效脱敏,达到数据安全、可信和可用的目标,已经成为了各行业数据产生者、使用者和管理者面临的巨大挑战。

针对异构大数据环境下不同应用场景下差异化的数据脱敏需求,本文提出了一种基于文本、音频、图片和数据库等多样化大数据的数据脱敏模型,该模型可实现不同应用场景下异构敏感数据的自动标注和分级,并通过抽取数据预脱敏处理和脱敏效果评估,实现了多应用场景下异构数据的高效脱敏。

1 数据脱敏基本框架

本文提出的数据脱敏方法基本框架主要包括 4 个模块,即源数据预处理模块、敏感数据标注模块、数据脱敏模块、脱敏数据输出及恢复模块,具体描述如下:

1) 源数据预处理模块。用户提交所需脱敏的源数据及应用场景。源数据的提交形式包括 txt、Word、Excel 等类型的文本数据,JPG、PNG 等类型的图片数据,MP3、WAV 等类型的音频数据,MySQL、HIVE 等数据库数据;应用场景包括金融、医疗、社交、教育、政府、零售等。该模块将原始数据及应用场景进行解析,并进行统一格式转换,将数据处理成下一模块能够识别的形式,用于敏感数据的识别和脱敏。

2) 敏感数据标注模块。该模块面向异构数据类型,采用机器学习方法和特定场景的敏感数据先验知识训练得到不同应用场景下的敏感数据识别模型^[16],实现敏感数据和非敏感数据的识别和标注,并且为敏感数据分级(等级为数值 1 ~ 10,等级越高,则该项数据越敏感)。

3) 数据脱敏模块。该模块首先以 10% 的比例随机抽取样例数据,判断其类型为文本、图片、音频或数据库内容,通过内容的不同,选择内置的多种脱敏方法实现预脱敏操作(文本脱敏包括 k -匿名、 l -多样性、 t -保密、差分隐私、对称加密、非对

称加密、保形加密和全同态加密等;图片脱敏包括人脸替换、高斯模糊等;音频脱敏包括空白音频替换等;数据库包括 AES 加密等)。然后,系统根据内置规则计算数据可用性、数据关联性、隐私保护度、时间复杂度、空间复杂度 5 个方面的内容,得到多种脱敏方法的脱敏效果评价结果,即多个 5 维向量,并根据综合评测推荐最合适的脱敏方法。最后,用户根据需求选择合适的脱敏方法,系统完成所有数据的脱敏操作。

4) 脱敏数据输出及恢复模块。将上述操作 1) 识别的敏感数据替换成脱敏后的数据,对授权用户可恢复成用户提交的原始数据格式,实现脱敏数据的授权访问。

2 异构大数据脱敏模型

图 1 为本文提出的面向异构大数据环境的数据脱敏模型,主要包括脱敏数据预处理、定制化脱敏策略、脱敏任务调度及脱敏数据恢复 4 部分。该模型集成了针对于文本、图片、音频和数据库 4 种类型数据的多种脱敏方法,对于数据库脱敏,运用 AES 对称加密算法对数据源进行脱敏;对于文本脱敏设计 8 种算法,分别是 k -匿名、 l -多样性、 t -保密、差分隐私、对称加密、非对称加密、保形加密和全同态加密;对于图片数据,针对人脸数据设计替换和高斯模糊方法 2 种脱敏方式;对于音频数据,设计用空白音频替换敏感音频。此外,本文提出一种脱敏效果评估方法,针对预脱敏的结果,从数据可用性、数据关联性、隐私保护度、时间复杂度、空间复杂度 5 个方面进行量化评估,得到各种方法预脱敏的评测结果,并且通过不同的权值和影响因子进行综合计算,为用户进行推荐,根据用户选择确定最终的脱敏方法。

2.1 脱敏数据预处理

本模块采用灵活配置的方式,支持用户实现对敏感数据的选择功能。在智能识别敏感数据的基础上,操作人员可以将敏感数据指定具体数据源、数据库、数据表及具体的属性字段上,以应对不同的业务需求。后续算法会根据用户的选择和定义将被指定为敏感的数据进行脱敏处理,而其他未指定的数据则保持不变。本模块由 2 个子模块构成:

1) 数据信息提取。采用人工配置或语句查询等方式,提取需进行脱敏的数据源名称、数据库名称列表、对应数据库中存储的数据库表列表、特定数据库表结构对应的数据字段及相应属性等信息。对于新接入及现有各类数据源,能够较为方

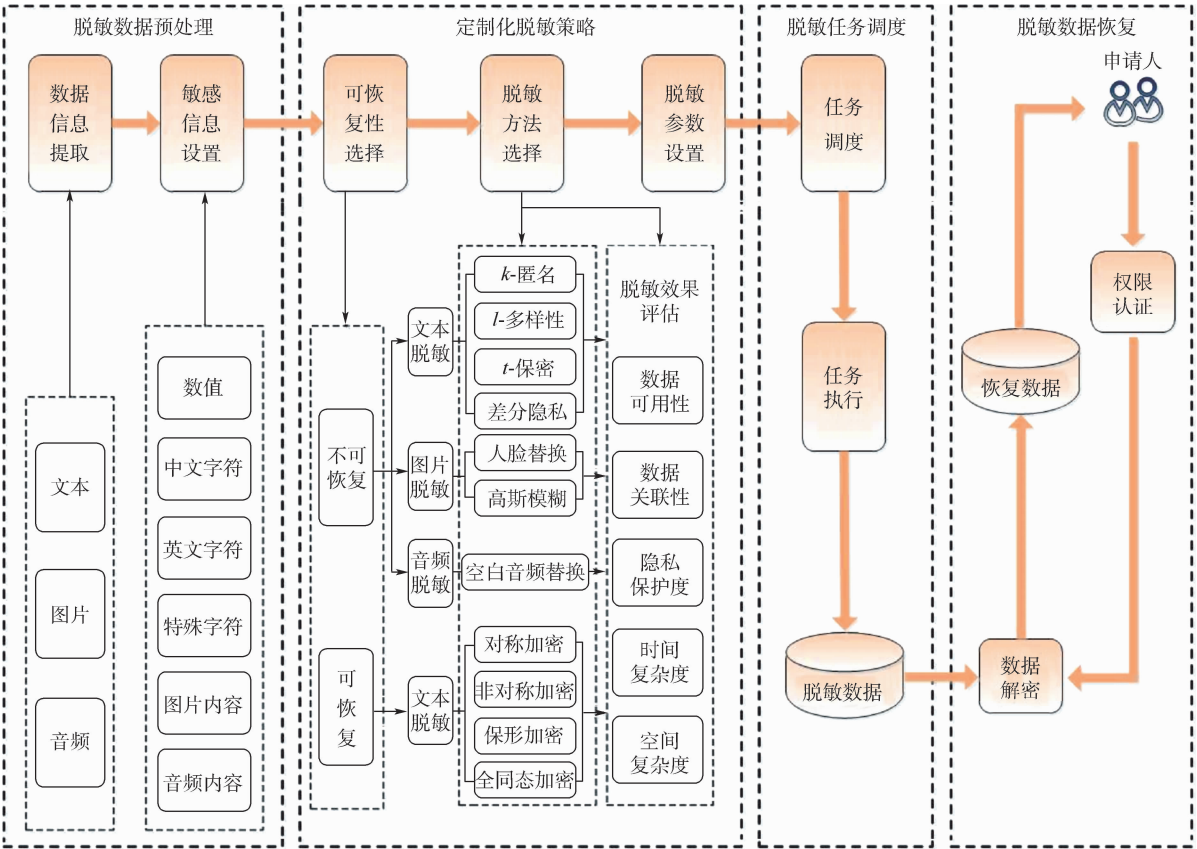


图 1 异构大数据脱敏模型

Fig. 1 Data masking model for heterogeneous big data

便总览其全部数据及不同数据表之间的关联信息,便于后续敏感数据的选择。

2) 敏感信息设置。用户按照法律法规或标准规范等要求对敏感数据预设分类,并依据不同应用场景需求构建原始敏感数据知识库和分级规则。对于待脱敏数据,依据其数据类型不同,分别采用自然语言处理和文本识别、多媒体内容理解和识别等技术,对待脱敏数据进行准实时处理,识别出敏感数据,同时设计人工反馈机制,可针对敏感数据识别结果进行修正,并逐步达到最优识别结果。此外,需明确各类敏感数据的具体数据类型,如针对文本数据是中文字符、英文字符、特殊字符等,以便后续最优脱敏策略的选择。

2.2 定制化脱敏策略

如图 1 所示,定制化脱敏策略模块主要包括可恢复性选择、脱敏方法选择、脱敏参数设置等。

1) 可恢复性选择。按照脱敏后的数据能否恢复到原始数据来划分,现有的脱敏方法可以分为可恢复与不可恢复两大类,以满足不同任务需求。可恢复方法主要以数据加密方法为主,在数据加密的过程中会同时生成相应的解密密钥等,加密后的数据可依据实际使用需求,通过解密密钥还原为原始数据;不可恢复数据脱敏方法在使

用匿名、替换等操作后,无法还原原始数据,因此在信息保留上存在一些损失。

2) 脱敏方法选择。根据用户在上一模块的选择,系统将分别提供不同的方法,这些方法各有优缺点,适用于不同的应用场景,用户可根据需求进行灵活选择。其中,主流不可恢复脱敏方法包括 k -匿名、 l -多样性、 t -保密、差分隐私;主流可恢复脱敏方法包括对称加密、非对称加密、保形加密、全同态加密等。其共同基本原理为:针对待脱敏数据,组合使用各类基础数据脱敏操作,以达到每种方法各自的脱敏规范。

针对不同数据类型,可使用的基础脱敏操作不同,具体的对应关系如表 1 所示。

表 1 不同数据类型的常用脱敏操作

Table 1 Commonly used data masking operation for different data types

操作类型	替换	改组	数字/日期差异	加密	删除	遮掩	平均值
数值	✓	✓	✓	✓	✓	✓	✓
中文字符	✓	✓		✓	✓	✓	
英文字符	✓	✓		✓	✓	✓	
特殊字符	✓	✓		✓	✓	✓	
图片	✓			✓			✓
音频	✓			✓			

3) 脱敏参数设置。一些方法具有可调节的参数。例如, k -匿名中的 k 可以看作是控制脱敏力度的指标。 k -匿名要求对于任意一行纪录, 其所属的相等集内纪录数量不小于 k , 即至少有 $k-1$ 条记录半标识列属性值与该条记录相同, 即增大 k 的值, 敏感数据会隐藏的更好, 但相应地需要操作的数据条数、数据长度也会增加, 因此导致更多的信息损失。用户可根据具体需求进行参数的设置, 以完成脱敏方法的针对性定制。

图 2 为基于该脱敏策略的定制化脱敏流程, 该过程中需要考虑的因素主要包括:

- 1) 数据可用性。即脱敏后的数据应能满足分析应用需求, 若脱敏后的数据无法用于目标分析及应用, 就不具有使用价值。在特定的应用场景中, 可能需要保留部分非关键信息(如身份证号码、手机号码的部分字段、数据的统计分析特征等)才能满足分析要求。
- 2) 数据关联性。对于结构化和半结构化的数据, 在同一数据表中某字段与另外字段有对应关系, 如果脱敏方法破坏了这种关系, 该字段的使用价值将不复存在。

3) 隐私保护度。数据中原本隐含的敏感信息在脱敏后被其他人获取的难易程度, 获取敏感信息的难度越高, 则隐私保护度越高。

4) 时间复杂度。即脱敏方法对数据进行脱敏所需要的时间, 在保证敏感数据满足保护条件的前提下, 所需计算时间越短越好。

5) 空间复杂度。即脱敏方法对数据进行脱敏所需要的存储和计算空间。

2.3 脱敏任务调度

- 1) 任务调度。由于本文模型是面向大数据应用环境的, 需要处理的数据量可能会非常大, 利用大数据计算环境执行分布式数据脱敏任务会极大提高任务的执行效率。任务调度模块通过对脱敏任务的数据量、各算法的执行效率、各节点的计算性能进行评估, 将总体脱敏任务分解为若干个子任务并分配给对应子节点执行。
- 2) 任务执行。根据任务调度结果, 将脱敏方法下发至各对应子节点, 子节点接受方法后开始对分配的敏感数据进行脱敏。脱敏完成后, 将结果返回并整合汇总形成最终脱敏后的数据结果。

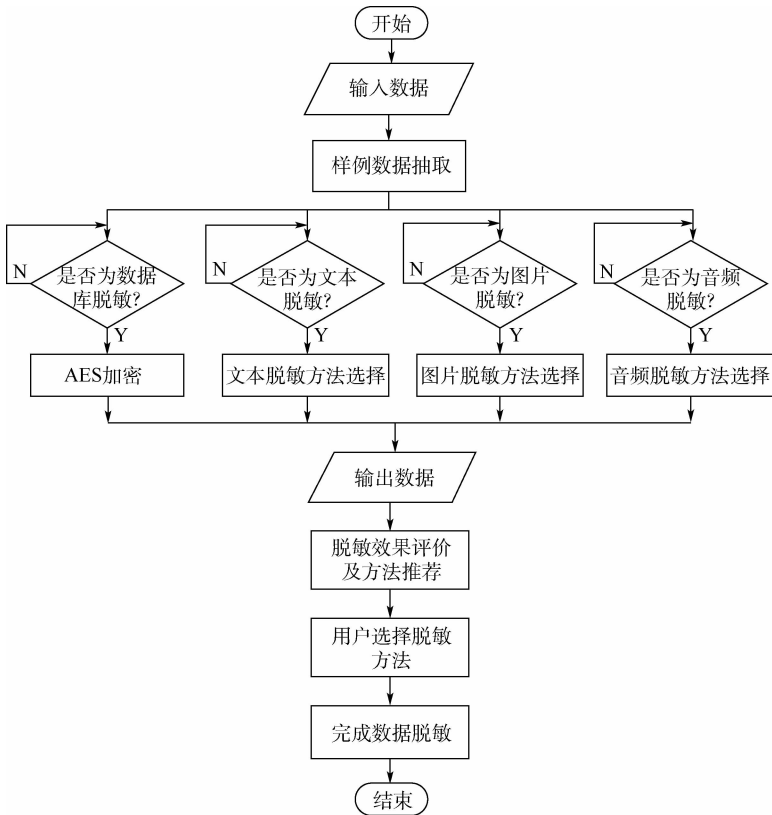


图 2 定制化数据脱敏流程

Fig. 2 Customized data masking process

2.4 脱敏数据恢复

由于敏感数据的重要性, 在实际应用中对部分脱敏数据的恢复时, 需要进行严格的权限管理。

申请人员需要先提出申请, 说明需要数据恢复的原因、需要恢复的具体数据项等信息, 并进行相应的权限认证。在通过权限认证模块审核后,

可在系统内执行相应的数据解密模块,并将数据恢复结果返回给特定申请用户。

2.5 脱敏模型安全性及可扩展性分析

本文所提异构大数据脱敏模型中 4 个关键模块具有较低的功能耦合度,又相互数据流关联实现了完整脱敏流程,从而使用该脱敏模型无需改变原有的业务数据逻辑,可确保数据整体安全性。此外,脱敏后数据恢复则需要严格权限认证才可获得解密密钥,满足了用户安全性需求。

本文提出的脱敏模型中定制化脱敏策略模块可提供灵活的脱敏规则配置及脱敏规则扩展,用户可以通过灵活调整配置参数,达到所设计的脱敏效果,实现了完全透明的、可扩展的数据脱敏处理能力。脱敏任务调度模块充分考虑了计算性能可扩展能力,通过采取多节点协作的分布式计算来提升计算效率,结合算法执行效率、单节点计算性能,将数据脱敏任务分解为若干个子任务后,分配给对应子节点进行计算,因此可方便增加子节点来扩展数据处理能力。

3 脱敏策略形式化描述

针对数据脱敏策略选择,通过层次分析法^[17]对用户数据脱敏需求和提供的脱敏方法匹配度进行了形式化描述和量化分析。该方法以脱敏数据的隐私保护度、数据可用性、数据关联性、时间复杂度、空间复杂度 5 项指标为用户选择脱敏方法的判断准则,主要包括 4 个步骤:①建立脱敏策略选择层次结构模型;②构造判断矩阵;③层次单排序及其一致性检验;④层次总排序及其一致性检验。

以面向文本数据的 8 种脱敏方法为例,对本文提出的脱敏策略执行过程进行描述。将决策的目标、考虑的因素(判断准则)和决策对象按相互关系分成目标层、准则层和方案层,形成层次化结构模型,如图 3 所示。

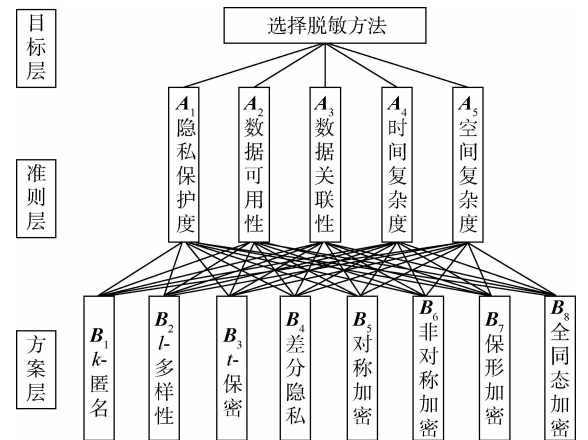


图 3 数据脱敏策略层次结构模型

根据用户选择脱敏方法的 5 项判断准则优先级要求,设定构造判断矩阵 $A = (a_{ij})$, a_{ij} 表示第 i 个与第 j 个元素相对上一层某个因素重要性的相对权重量化值,并按下述标度进行赋值: $a_{ij} = 1$, 元素 i 与元素 j 对上一层因素的重要性相同; $a_{ij} = 3$, 元素 i 比元素 j 略重要; $a_{ij} = 5$, 元素 i 比元素 j 重要; $a_{ij} = 7$, 元素 i 比元素 j 重要得多; $a_{ij} = 9$, 元素 i 比元素 j 极其重要; $a_{ij} = 2n$, $n = 1, 2, 3, 4$, 元素 i 比元素 j 的重要性介于 $a_{ij} = 2n - 1$ 与 $a_{ij} = 2n + 1$ 之间。

为分析一致性,先计算描述一个成对比较矩阵 A ($n > 1$ 阶方阵)不一致程度的指标 CI:

$$CI = \frac{\lambda_{\max}(A) - n}{n - 1} \tag{1}$$

式中: $\lambda_{\max}$ 为判断矩阵 A 的最大特征值; n 为矩阵 A 的阶数。

为衡量 CI 的大小,引入随机一致性指标 RI,如表 2 所示。

表 2 随机一致性指标 RI 的数值

Table 2 Values of random consistency indicator RI									
n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

定义一致性比率为 $CR = CI/RI$,且当一致性比率 $CR < 0.1$ 时,判定 A 能够通过一致性检验,否则需重新构造对比矩阵 A 。计算矩阵 A 的特征向量得到 $U = (u_1, u_2, u_3, u_4, u_5)$, u_i 表示每项脱敏效果评价指标对应最终脱敏方法选择的权重。

在上述基础上,计算方案层对最终目标层的总排序权向量,先需根据各脱敏方法在每个准则层指标上的比较,得到每个指标的判断矩阵。根据候选脱敏方法每项指标的对比,得到方案层的多个判断矩阵 B_i ,对每一个判断矩阵进行一致性检验和计算求权向量。最终将方案层与之前准则层的对比矩阵进行权向量组合,即可得到方案层对目标层的组合权向量,取其中权重最高者对应的脱敏方法,即为最适合用户需求的脱敏方案。

4 应用场景验证

基于以上构建的异构大数据脱敏模型,本文在 2 类大数据脱敏应用场景中进行了验证,具体脱敏验证流程分别描述如下。

4.1 银行客户贷款信息数据脱敏

图 4 为某银行客户贷款信息数据脱敏过程,其主要步骤包括:

Fig. 3 Hierarchical model of data masking strategy

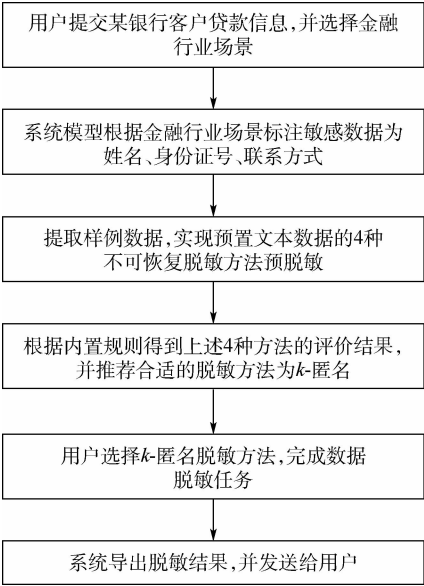


图4 某银行客户贷款信息数据脱敏过程

Fig. 4 Data masking process for bank customer loan information

1) 用户提交某银行客户贷款信息,该提交信息为 Excel 表格,包括姓名、性别、身份证号、籍贯、贷款金额、贷款日期、联系方式等内容,并选择金融行业场景。

2) 根据金融行业场景标注的非敏感数据为性别、籍贯、贷款金额、贷款日期等,敏感数据为姓名、身份证号、联系方式。

3) 以 10% 的比例随机抽取样例数据,判断其为文本数据,用户选择 4 种不可恢复文本脱敏方法(k -匿名、 l -多样性、 t -保密、差分隐私)实现预脱敏操作。通过分析用户需求,对准则层指标的优先级要求为:时间复杂度 > 空间复杂度 = 隐私保护度 > 数据可用性 > 数据关联性。

4) 根据用户设定的指标优先级,采用层次分析法构建判断矩阵 A 如下:

$$A = \begin{bmatrix} 1 & 2 & 3 & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & 3 & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 1 & \frac{1}{8} & \frac{1}{3} \\ 2 & 3 & 8 & 1 & 2 \\ 1 & 2 & 3 & \frac{1}{2} & 1 \end{bmatrix}$$

计算判断矩阵 A 的最大特征值 $\lambda_{\max} = 5.053$, $CI = 0.013\ 25$, $RI = 1.12$, $CR = 0.012 < 0.1$ 说明对比矩阵具有良好的一致性。该矩阵对应的特征向量 $U = (-0.395\ 1, 0.508\ 7, 0.508\ 7, 0.707\ 1, -0.041\ 7)$, 对该特征向量进行标准化,得到准则

层对比矩阵的权向量 $U_0 = (0.182\ 8, 0.235\ 4, 0.235\ 4, 0.327\ 1, 0.019\ 3)$ 。

构建方案层每一种脱敏方法对于准则层隐私保护度、数据可用性、数据关联性、时间复杂度、空间复杂度 5 项指标的判断矩阵,分别为 B_1 、 B_2 、 B_3 、 B_4 、 B_5 :

$$B_1 = \begin{bmatrix} 1 & 2 & 3 & 5 \\ \frac{1}{2} & 1 & 2 & 4 \\ \frac{1}{3} & \frac{1}{2} & 1 & 2 \\ \frac{1}{5} & \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} 1 & 4 & 3 & 2 \\ \frac{1}{4} & 1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 3 & 1 & \frac{1}{2} \\ \frac{1}{2} & 3 & 2 & 1 \end{bmatrix}$$

$$B_3 = \begin{bmatrix} 1 & 4 & 3 & 2 \\ \frac{1}{4} & 1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 3 & 1 & \frac{1}{2} \\ \frac{1}{2} & 3 & 2 & 1 \end{bmatrix}$$

$$B_4 = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{5} & \frac{1}{7} \\ 4 & 1 & \frac{1}{2} & \frac{1}{4} \\ 5 & 2 & 1 & \frac{1}{2} \\ 7 & 4 & 2 & 1 \end{bmatrix}$$

$$B_5 = \begin{bmatrix} 1 & 4 & 2 & 1 \\ \frac{1}{4} & 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 2 & 1 & \frac{1}{2} \\ 1 & 4 & 2 & 1 \end{bmatrix}$$

经过对 B_1 、 B_2 、 B_3 、 B_4 、 B_5 进行一致性检验,其一致性比率分别为 0.005 8、0.024 3、0.024 3、0.018 0,均符合一致性要求。归一化后的特征向量分别为: $U_1 = (0.320\ 6, 0.339\ 7, 0.339\ 7, 0)$, $U_2 = (-0.257\ 2, 0.244\ 5, 0.249\ 2, 0.249\ 2)$, $U_3 = (-0.257\ 2, 0.244\ 5, 0.249\ 2, 0.249\ 2)$, $U_4 = (0.496\ 9, -0.251\ 6, -0.251\ 6, 0)$, $U_5 = (0.106\ 8, 0.225\ 6, -0.421\ 3, -0.244\ 9)$ 。

隐私保护度、数据可用性、数据关联性、时间

复杂度、空间复杂度分别为 $Z_1、Z_2、Z_3、Z_4、Z_5$ 。差分隐私、 t -保密、 l -多样性、 k -匿名分别为 $F_1、F_2、F_3、F_4$,则 $Z_1、Z_2、Z_3、Z_4、Z_5$ 脱敏方法的选择权重为 $(0.182\ 8,0.235\ 4,0.235\ 4,0.327\ 1,0.019\ 3)$, $F_1、F_2、F_3、F_4$ 对 $Z_1、Z_2、Z_3、Z_4、Z_5$ 的权重分别为: $(0.320\ 6,0.339\ 7,0.339\ 7,0)$, $(-0.257\ 2,0.244\ 5,0.249\ 2,0.249\ 2)$, $(-0.257\ 2,0.244\ 5,0.249\ 2,0.249\ 2)$, $(0.496\ 9,-0.251\ 6,-0.251\ 6,0)$, $(0.106\ 8,0.225\ 6,-0.421\ 3,-0.244\ 9)$ 。因此, F_1 对脱敏方法选择的权重为: $0.320\ 6 \times 0.182\ 8 - 0.257\ 2 \times 0.235\ 4 - 0.257\ 2 \times 0.235\ 4 + 0.496\ 9 \times 0.327\ 1 + 0.106\ 8 \times 0.019\ 3 = 0.102\ 1$, F_2 对脱敏方法选择的权重为: $0.339\ 7 \times 0.182\ 8 + 0.244\ 5 \times 0.235\ 4 + 0.244\ 5 \times 0.235\ 4 - 0.251\ 6 \times 0.327\ 1 + 0.225\ 6 \times 0.019\ 3 = 0.099\ 2$, F_3 对脱敏方法选择的权重为: $0.339\ 7 \times 0.182\ 8 + 0.249\ 2 \times 0.235\ 4 + 0.249\ 2 \times 0.235\ 4 - 0.251\ 6 \times 0.327\ 1 - 0.421\ 3 \times 0.019\ 3 = 0.089\ 0$, F_4 对脱敏方法选择的权重为: $0 \times 0.182\ 8 + 0.249\ 2 \times 0.235\ 4 + 0.249\ 2 \times 0.235\ 4 + 0 \times 0.327\ 1 - 0.244\ 9 \times 0.019\ 3 = 0.112\ 6$ 。 F_4 方法对脱敏方法选择的权重最大,因此最终选择的脱敏方法为 k -匿名。

- 5) 系统采用 k -匿名方法完成数据脱敏任务。
- 6) 系统导出脱敏结果,将姓名、身份证号、联系方式替换成脱敏后的内容,并恢复成用户提交的 Excel 表格格式,发送给用户。

4.2 社交网络图片数据脱敏

图 5 为社交网络图片数据脱敏过程,其主要步骤包括:

- 1) 用户提交社交网络图片集,该提交信息为包含多个 JPG 文件的文件夹,图片内容涵盖人脸、风景、动物、美食、汽车,并选择社交应用场景。
- 2) 根据金融行业场景标注的非敏感数据为风景、动物、美食、汽车,敏感数据为人脸,敏感层级为 9。
- 3) 以 10% 的比例随机抽取样例数据,判断其为图片内容,通过预置的 2 种图片脱敏方法(换脸、高斯模糊)实现预脱敏操作。
- 4) 采用层次分析法,得到上述 2 种方法的评价结果,并根据综合评测推荐合适的脱敏方法为换脸方法。
- 5) 若考虑到换脸后的图片难以满足业务需求,用户实际选择的脱敏方法为高斯模糊方法,系统完成数据脱敏任务。
- 6) 系统导出脱敏结果,将原人脸替换成高斯模糊后的人脸,并保存为 JPG 文件,发送给用户。

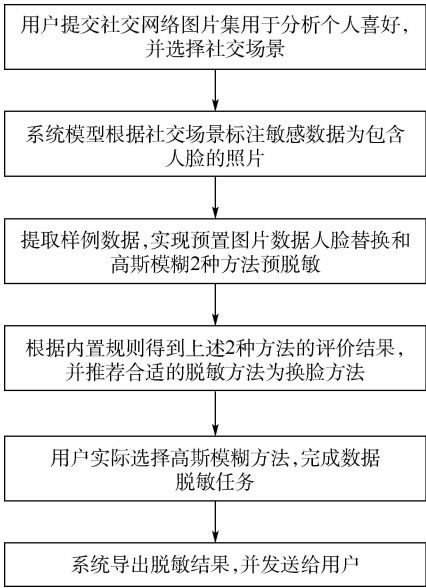


图 5 社交网络图片数据脱敏过程
Fig. 5 Data masking process for social network pictures

5 结 论

本文主要针对当前异构大数据中敏感信息的精准定向、高效脱敏等研究难点,提出一种在异构大数据环境下,基于文本、图片、音频和数据库等异构数据的脱敏模型,并在实际应用场景中进行了验证,得到结论如下:

- 1) 通过脱敏数据预处理,实现不同应用场景下敏感数据的自动标注和分级,有效降低实际应用场景脱敏处理复杂度,更有利于脱敏数据和脱敏策略的选择。
- 2) 数据脱敏模型实现定制化脱敏策略,并利用数据预脱敏处理方法,从数据可用性、数据关联性、隐私保护度、时间和空间复杂度等 5 个维度进行脱敏效果评价。
- 3) 通过银行客户贷款信息数据、社交网络图片数据脱敏应用场景进行验证,表明本文提出的异构大数据脱敏模型能够实现不同应用场景下异构敏感数据的高效脱敏。

参考文献 (References)

[1] SWEENEY L. k -anonymity: A model for protecting privacy[J]. Fuzziness and Knowledge-based Systems, 2002, 10 (5) : 557-570.

[2] RADHAKRISHNAN R, KHARRAZI M, MEMON N. Data masking: A new approach for steganography[J]. Journal of VL-SI Signal Processing Systems for Signal, Image and Video Technology, 2005, 41 (3) : 293-303.

[3] RAVIKUMAR G K, MANJUNATH T N, RAVINDRA S, et al. A survey on recent trends, process and development in data masking for testing[J]. International Journal of Computer Science,

2011,8(2):535-544.

[4] VICTOR N,LOPEZ D,ABAWAJY J H. Privacy models for big data;A survey [J]. International Journal of Big Data Intelligence,2016,3(1):61-75.

[5] VADREVU P K,ADUSUMALLI S K,MANGALAMPLLI V K. Survey;Privacy preserving data publication in the age of big data in IoT era [J]. International Journal of Engineering, Science and Mathematics,2017,6(8):938-944.

[6] 陈天堂,陈剑锋. 大数据环境下的智能数据脱敏系统 [J]. 通信技术,2016,49(7):915-922.

CHEN T Y,CHEN J F. Intelligent data masking system for big data productive environment [J]. Communications Technology, 2016,49(7):915-922(in Chinese).

[7] MACHANAVAJJHALA A,GEHRKE J,KIFER D, et al. *l*-diversity;Privacy beyond *k*-anonymity [C] // IEEE 22nd International Conference on Data Engineering. Piscataway: IEEE Press,2006:24.

[8] LI N,LI T,VENKATASUBRAMANIAN S. *t*-closeness; Privacy beyond *k*-anonymity and *l*-diversity [C] // IEEE 23rd International Conference on Data Engineering. Piscataway: IEEE Press,2007:106-115.

[9] SARADA G,ABITHA N,MANIKANDAN G, et al. A few new approaches for data masking [C] // International Conference on Circuits, Power and Computing Technologies. Piscataway: IEEE Press,2015:15295632.

[10] GUJJARY V A,SAXENA A. A neural network approach for data masking [J]. Neurocomputing,2011,74(9):1497-1501.

[11] ZHOU Y,LOUIS T A. A smoothing approach for masking spatial data [J]. Annals of Applied Statistics,2010,4(3):1451-1475.

[12] 吴克河,朱海,李为,等. 基于敏感信息度量的 *t*-保密脱敏技术改良 [J]. 信息技术,2019(11):5-9.

WU K H,ZHU H,LI W, et al. An improvement of *t*-closeness technology based on sensitive information metric [J]. Information Technology,2019(11):5-9(in Chinese).

[13] SANTOS R J, BERNARDINO J, VIEIRA M. A data masking technique for data warehouses [C] // Proceedings of the 15th Symposium on International Database Engineering & Applications,2011:61-69.

[14] 张琦颖. 大数据脱敏系统的设计与实现 [D]. 北京:北京邮电大学,2018:19-33.

ZHANG Q Y. The design and implementation of big data anonymity system [D]. Beijing:Beijing University of Posts and Telecommunications,2018:19-33(in Chinese).

[15] 邵华西. 基于 *T*-Closeness 的大数据脱敏系统的设计与实现 [D]. 北京:北京邮电大学,2019:44-52.

SHAO H X. Design and implementation of *T*-Closeness based big data anonymization system [D]. Beijing:Beijing University of Posts and Telecommunications,2019:44-52(in Chinese).

[16] 王鑫,王电钢,母继元,等. 基于机器学习的数据脱敏系统研究与设计 [J]. 电力信息与通信技术,2018,16(1):33-38.

WANG X,WANG D G,MU J Y, et al. Research and implementation of data masking system based on machine learning [J]. Electric Power ICT,2018,16(1):33-38(in Chinese).

[17] 邓雪,李家铭,曾浩健,等. 层次分析方法权重计算方法分析及其应用研究 [J]. 数学的实践与认知,2012,42(7):93-100.

DENG X,LI J M,ZENG H J, et al. Research on computation methods of AHP weight vector and its applications [J]. Mathematics in Practice and Theory,2012,42(7):93-100(in Chinese).

Data masking model for heterogeneous big data environment

TONG Lingling, LI Pengxiao^{*}, DUAN Dongsheng, REN Boya, LI Yangxi

(National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Due to the variety of data types and desensitization demand in different scenarios, traditional data masking methods cannot meet the user privacy protection requirements in the environment of big data. How to realize the accurate pointing and efficient desensitization of heterogeneous big data for data security, trust and availability, has become the key in this area. In this paper, we propose a data masking model for heterogeneous big data applications, such as texts, images, voices and databases, and four key modules are presented in our model. First, the sensitive data automatic identification and classification in different applications are realized in different application scenarios by desensitization data preprocessing. Second, with data pre-masking method, the data masking evaluation is implemented in five dimensions, including data availability, data relevance, degree of privacy protection, and time and space complexity, to construct the customized desensitization strategy. Finally, after task scheduling, the allocation and execution of the data masking tasks are performed, and the masking data recovery can also be partially supported. Two typical data masking applications are verified and analyzed based on the proposed heterogeneous big data masking model, indicating that effective desensitization can be achieved in different application scenarios.

Keywords: heterogeneous big data; sensitive data automatic annotation; data masking; data masking evaluation; machine learning

Received: 2020-08-09; **Accepted:** 2020-09-05; **Published online:** 2020-11-06 14:49
URL: kns.cnki.net/kcms/detail/11.2625.V.20201105.1659.005.html
Foundation items: National Natural Science Foundation of China (U1936110, U1836111)
*** Corresponding author.** E-mail: lp@cert.org.cn

http://bhxb.buaa.edu.cn jhbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0539

基于异质信息网络的恶意代码检测

刘亚姝¹, 侯跃然², 严寒冰^{3,*}

(1. 北京建筑大学 电气与信息工程学院, 北京 100044; 2. 北京邮电大学 网络技术研究院, 北京 100876;
3. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 恶意代码对网络安全、信息安全造成了严重威胁。如何快速检测恶意代码, 阻止和降低恶意代码产生的危害一直是亟需解决的问题。通过获取恶意应用的动态信息、构造异质信息网络(HIN), 提出了描述恶意代码动态特征的方法, 实现了恶意代码检测与分类。构建了 FILE、API、DLL 三类对象的 4 种元图, 刻画了恶意代码 HIN 的网络模式。经过改进的随机游走策略, 尽可能多地获取元图中对象节点的上下文信息, 将其作为连续词包(CBOW)模型的输入, 从而得到词向量的网络嵌入。通过投票方法改进主角度分析模型, 得到多元图特征融合的分类结果。在仅可获得有限信息的情况下, 大大提高了基于单元图特征的恶意样本分类准确率。

关 键 词: 恶意代码; 异质信息网络(HIN); 随机游走; 连续词包(CBOW); 元图
中图分类号: TP393

文献标志码: A **文章编号:** 1001-5965(2022)02-0258-08

恶意代码可以定义为任意对用户、计算机或网络做出危害的软件^[1]。一般来讲, 恶意代码包括病毒、蠕虫、后门、漏洞攻击程序、特洛伊木马、rootkit 等。随着技术的发展, 恶意代码的产生越来越简单, 恶意代码的变体也越来越多, 对网络安全和信息安全造成了严重的威胁。

恶意代码的检测可以分为静态分析和动态分析两大类。动态分析需要在沙箱中或者虚拟机里实际运行可执行文件, 分析执行程序的行为来识别各种潜在的危害行为和攻击行为, 以便识别代码的恶意性。动态分析是恶意代码分析的一个重要研究方向, 相对静态分析方法而言, 动态分析更常用于工业应用中。

本文通过在沙盒中观察样本的动态执行过程, 提取动态 API 的访问序列、调用的 DLL 等信息构建异质信息网络(heterogeneous information

network, HIN)。在所获得的节点类型不充足的情况下, 改进随机游走策略, 选取合适的方法描述元图特征, 从而实现恶意样本的检测和分类。

图是一种直观地表现出对象之间关系的工具。通常用有向图描述对象之间的联系, 这样的有向图也称为信息网络。图中顶点表示对象、边表示对象间联系。如果信息网络中描述的对象类型是不同的, 也就意味着图中有多种类型的数据及多种关系, 则被称为异质信息网络, 否则称为同质信息网络^[2-3]。与同质信息网络相比, 异质信息网络更能反映真实条件下对象之间的关系, 具有更丰富的结构和语义信息。常见的异质信息网络有社交网络、科技文献、医疗系统等。

由于异质信息网络的研究对象类型、链接关系多样化, 传统的图方法并不适用。2011 年, Sun 等^[4]提出了元路径(meta path)的概念, 给出了异

收稿日期: 2020-09-23; 录用日期: 2020-12-18; 网络出版时间: 2021-04-27 11:10
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210426.1819.003.html
基金项目: 国家重点研发计划(2018YFB0803604, 2018YFB0804704); 国家自然科学基金(U1736218); 北京建筑大学市属高校基本科研业务费专项资金(X20152)
* 通信作者. E-mail: yhb@cert.org.cn

引用格式: 刘亚姝, 侯跃然, 严寒冰. 基于异质信息网络的恶意代码检测[J]. 北京航空航天大学学报, 2022, 48(2): 258-265.
LIU Y S, HOU Y R, YAN H B. Malicious code detection based on heterogeneous information network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 258-265 (in Chinese).

质网络的分析工具,通过对称元路径抽取 2 个节点之间的链接关系,同时定义了一种名为 PathSim 的相似度度量方法,用于在异构网络中寻找对等对象。2012 年,Sun 等^[5]提出了元路径的聚类方法——PathSelClus。异质信息网络中一个很重要的任务是将元路径提取的分散语义信息进行聚合。PathSelClus 采用在用户提供一个小的对象种子集合的指导下(称为用户指导)实现元路径聚类。PathSelClus 方法可以自动学习和设置元路径聚集时的权重问题,与其他方法相比具有聚类结果稳定、隐含语义信息的效果。2014 年,Shi 等^[6]提出了在异质信息网络中用统一框架判断相同类型节点和不同类型节点的相关性方法——HeteSim。异质信息网络中的元图会存在奇数路径(odd path),也就是路径不对称及任意路径(arbitrary path)的问题。HeteSim 通过约束路径搜索的方法将奇数路径转换为等长路径(even path),这样 2 个节点之间的相关性概率通过路径上的所有概率之和获得,从而能够计算在元图中无路径到达的 2 个节点的相关性。2015 年,Cao 等^[7]提出了 HCLP 方法用于节点之间多种类型的链接关系的预测,给出了相似链接规则及不同类型节点间相关性度量方法(RM)以便计算不同类型节点间的存在概率。2016 年,Huang 等^[8]建议使用元图/元结构(meta graph/meta structure)度量对象之间的相似程度。元图是一种有向无环图,可以描述 2 个对象之间的复杂关系。例如,文献[8]中指出 KDD 上的 2 篇论文具有相同类型的主题和相关的作者,这样就改变了原有元路径方法,必须是顺序结构的要求,并给出了基于元图的 2 个节点之间的相关性判断策略。

2014 年,Tamersoy 等^[9]基于置信网络传播算法,通过未知文件与已有标签的文件的关系实现无标签文件恶意性的判断。2015 年,Chen 等^[10]将同质信息网络用于恶意代码文件关系网络的构建。2018 年,Fan 等^[11]将异质信息网络应用到恶意代码检测中,通过多种信息——压缩包、API、DLL 及上报恶意文件所在的机器或位置等信息构建异质信息网络,获得其语义信息,通过不同的元图获得多种类型的特征描述,并通过主角度分析将特征向量映射到低维空间。该成果在 Comodo 反病毒产品中得到了应用。

本文将异质信息网络应用于恶意代码动态分析中,通过改进 Metagraph2vec 游走方式及网络嵌入方法,以获得样本特征描述,并将该特征用于恶意代码分类。实验结果证明,本文方法在仅可获

得有限信息的情况下,可以大大提高恶意代码检测、分类的性能。

1 相关工作

目前,异质信息网络最新分析工具是元路径和元图,下面给出元路径和元图的详细说明。

1.1 元路径与元图

2011 年,Sun 等^[4]提出了元路径的概念,用于异质信息网络中网络模式的特征描述。下面给出相关的符号定义。一个异质信息网络定义为 $G = (V, \varepsilon)$

其网络模式为

$$T_c = (A, R)$$

式中: A 为有向图 T_c 的顶点; R 为有向图的边。则有对象类型的映射 $\tau: V \rightarrow A$,链接的映射 $\Phi: \varepsilon \rightarrow R$ 。

元路径是网络模式 T_c 上的路径,记作 P ,形式定义为 $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$,定义复合关系 $R = R_1 \circ R_2 \circ \cdots \circ R_{l+1}$ 。

图 1 给出了 P_1 和 P_2 路径(S 为元图结构)。可以通过统计 2 个对象节点 x 和 y 的路径数量、随机游走等方法计算对象节点的相似度 $s(x, y)$,从而构建节点的特征向量。元路径直接反映了对象节点之间的单一链接关系,如果对象间关系复杂,则需要处理成多条元路径。如图 1 中的有向图 S 所示,需要处理成 P_1 和 P_2 路径并做线性组合。但是,元路径这样的处理方法无法约束节点之间的共享关系,会造成共享信息丢失的问题。为此,Huang 等^[8]提出了元图的概念,来刻画节点间的信息共享关系。元图相关性度量的方法有元图计数、结构约束的子图扩展等。

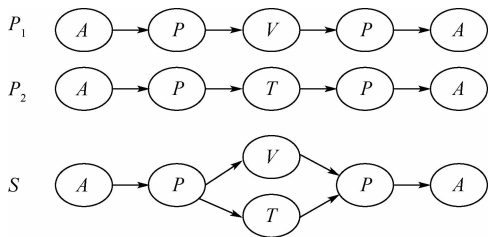


图 1 元路径与元图示例

Fig. 1 Examples of meta path and meta graph

1.2 Metapath2vec

2014 年,Perozzi 等^[12]提出了 DeepWalk 方法用于社交网络的特征表示。DeepWalk 使用随机游走(random walk)的方式在网络中查找当前节点的“上下文”节点,体现了节点与节点的共现关系。随机游走从局部一定程度上保持了节点与其相邻节点之间的连接性,即网络结构信息。然而

对于异质信息网络来说,由于节点与连接节点的异质性的存在,异质信息网络嵌入最大的难点在于如何有效地在多种类型节点之间保存“节点上下文”的信息。

2017 年,Yu 等^[13]提出了异质网络嵌入模型 Metapath2vec。Metapath2vec 是基于预先指定的元路径进行随机游走的。通过“有指导的”元路径游走,从而能够保持“节点上下文”的概念。

Metagraph2vec 是在 Metapath2vec 的基础上提出来的。区别之处是:Metagraph2vec 同时有 2 个随机游走的路线,如果初始给定的 2 个游走的起始点一致,Metagraph2vec 退化为 Metapath2vec。若元图 M 的基本模式如图 2 所示,则可以有 PVP 和 PTP 两个游走路径。

采用 skip-gram 建模,并分别做归一化得到节点的表示。在多元图上会得到同一个节点的不同表示,则需要通过权重的设置将这些节点表示融合起来。

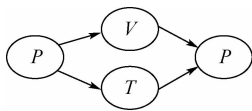


图 2 元图 M 示例

Fig. 2 Examples of meta graph M

2 恶意样本的异质信息网络构建

应用程序对系统资源访问、注册表访问、活跃的进程、被进程载入的 DLL、进程属性等,都是其是否具有恶意性的判断标志。而这些信息仅仅依靠对样本的静态分析是很难获得的,有很多恶意的特征只有在执行过程中才会动态展现。

本文主要从动态分析角度提取样本的特征,并基于异质信息网络构建多种类型对象之间的关系,采用改进的随机游走策略获得特征向量并实现样本恶意性的检测。

2.1 恶意代码网络模式的定义

Fan 等^[11]刻画的恶意代码异质信息网络中的信息包括:上报样本所在的主机(M)、样本文件(F)、文件所属的压缩包(Z)、动态调用的 API(A)、动态调用的 DLL(D)。分析这些类别信息之间存在的关系是有利于恶意代码检测的。例如,在 2 个 PE 文件中,如果其具有相同的压缩包、有属于同一个 DLL 的 API,那么这 2 个 PE 文件具有很高的概率同属于一个类别。图 3 刻画了以上述类别为顶点的恶意代码网络模式。

Fan 等^[11]研究中使用的数据集来源于 Como-

do 反病毒平台,可以获得图 3 中全部类型的节点数据。但是,通常在研究中很难获取机器信息、压缩包信息等。为此,如何在有限信息的情况下获得更好的异质信息网络的描述是本文的研究重点。

在恶意代码动态分析中,API(A)、DLL(D)是最常能够获取到的信息。本文在构建异质信息网络时,不仅考虑了样本的内容——API、DLL,也考虑了样本文件(F)与样本之间的关系。图 3 中虚线框内部分为本文构建的异质信息网络。

根据图 3 中虚线框内的网络模式可以得到图 4 所示的 4 种元图结构。

图 4 所示的元图中,节点之间涉及到的关系如下:

1) 关系 R_1 。被调用 API 是否属于同一个文件。用矩阵 I 描述,这里 $s_{ij} \in \{0, 1, \dots, n\}$ 表示文件 i 是否包含了第 j 个 API 及包含的次数。

2) 关系 R_2 。被提取的 API 是否属于同一个 DLL(如“WriteFile”与“CreateFileA”都属于“KERNEL32.DLL”)。用矩阵 B 描述,这里 $b_{ij} \in \{0, 1\}$ 描述第 i 个 API 是否在第 j 个 DLL 中。

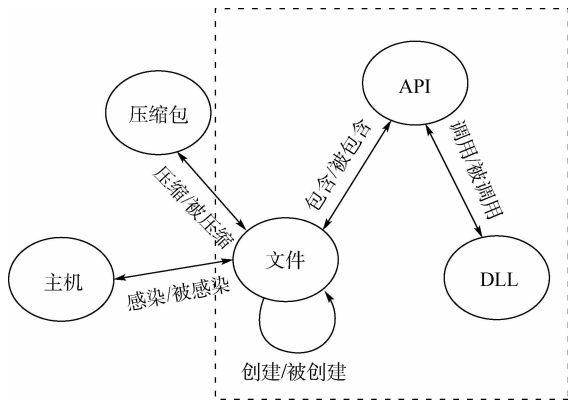


图 3 恶意代码网络模式

Fig. 3 Network schema of malicious code

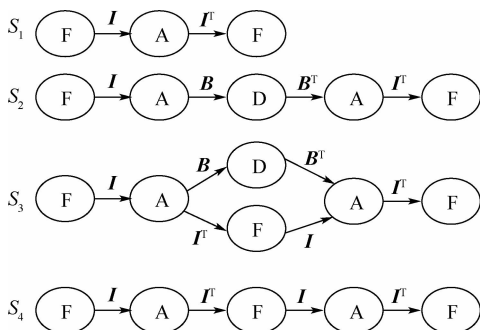


图 4 四种元图

Fig. 4 Four types of meta graph

2.2 改进的 Metagraph2vec

从图 4 中给出的 4 个元图结构上可以看到,在可获得对象类别有限的情况下,通过对象类别之间构造多种链接结构会获得更多的信息。为了

获得恶意代码异质信息网络的描述,本文采用了改进的随机游走策略。

根据元路径的对称性,在实际游走的过程中做了加速处理;限定的、多次随机游走,以获得尽可能多的上下文信息。将随机游走获取的节点上下文的信息作为 CBOW 输入,提取节点对象的特征向量。

2.2.1 改进的随机游走策略

随机游走模型用来描述不稳定的移动。以图 4 中 S_1 为例说明如何通过随机游走获取样本文件的环境信息(上下文信息)。为了在大量数据时游走算法依然保持灵敏,本文采取如下加速策略:

- 1) 减少大量相关但帮助性不强的信息。采用 TF-IDF 将普遍出现的关系信息去除。
- 2) 由于所有的元图都是在 2 个样本文件之间对称,仅仅取其中一半的节点进行随机游走,就可以刻画在这个元图下样本文件的周围环境信息。

关系 R_1 中反映的是样本文件与其包含的 API 及 API 出现的次数。则可根据样本文件中所包含的 API 数量确定随机游走的概率。通过游走可获得 $F \rightarrow A \rightarrow F$ 之间的关系。

图 5 为关系 R_1 的矩阵 I 的信息示例。 F_1 的 API 信息在第 2 行,先从行的角度,随机抽取 API 的信息,假设抽取到第 2 列数据,即 A_1 ,然后沿着第 2 列继续抽样,抽取到某一行,即得到 F_i 。这样的游走是在元图 S_1 指导下的游走,获得 $F \rightarrow A \rightarrow F$ 之间的关系。若从 F_i 出发继续游走,获取 A_j 信息,从 A_j 列游走获取 $F_k, \dots$,这样的游走是在元图 S_4 的指导下的游走。以上游走需要进行多次,以便获得更为丰富的信息。

游走获得的数据具有实际的物理含义,行表示的是每个样本文件中出现的 API 信息。根据 F_i 行游走得到的 A_i 信息,即得到样本文件中包含 A_i (API) 的次数,按照图 4 所示的 4 个元图,根据元图结构在关系 $R_1、R_2$ 上随机游走得到 A_i 信息,即得到样本文件中包含 A_i (API) 的次数,根据 A_i 获得该 API 在其他样本文件中出现的次数。

按照图 4 所示的 4 个元图,根据元图结构在

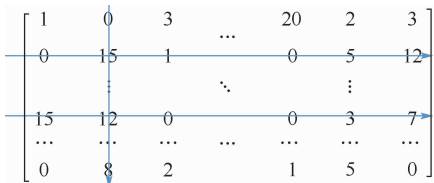


图 5 关系 R_1 的矩阵 I 示例

Fig.5 Example of matrix I on relationship R_1

关系 $R_1、R_2$ 上随机游走获得相对应的上下文信息,以此作为各个元图的词向量模型。

2.2.2 CBOW 模型与 skip-gram 模型

CBOW 模型与 skip-gram 模型都包含 3 层结构——输入层、投影层和输出层^[14-15]。

skip-gram 模型的训练方法是:中心词和周围某个词成对进行训练,中心词做输入、周围某个词做输出,描述中心词对周围词的推断。因此,可以将 skip-gram 输出词向量理解为对输入词典的降维。而 CBOW 模型的训练,是将中心词周围多个词和中心词进行训练,周围多个词做输入,中心词做输出,描述通过周围词推测中心词的过程。

本文采用 CBOW 模型进行词向量的网络嵌入,原因为:①图 4 所示的 4 个元图是从不同角度获得样本文件与周围 API、DLL 异构节点的分布关系。②使用 CBOW 模型可以直接获得样本的特征。

3 实验与分析

本文实验采用了 CNCERT 实验室提供的数据集,该数据集共有 10 个家族,每个家族 1 500 个恶意样本,共 15 000 个恶意代码样本。由于动态分析需要实际执行每一个样本,非常耗时。在实验中随机选择 3 个家族,下面以某次随机选取的 3 个家族 Allapple、Virut、Agent 为例完成实验结果分析和评价。

实验中为了获得样本的动态信息,需要在 Cuckoo 沙箱内运行样本。Cuckoo 沙箱是一款著名的开源沙箱系统,已被业界广泛使用。在 Cuckoo 沙箱内运行这 3 个家族 4 500 个样本,共获取 3 948 个有效样本的信息。

3.1 异质信息网络构建与信息表示

通过解析 Cuckoo 报告文件——“report. json”包含的字典中“apistats”子项,获得各个进程下的 API 统计信息;解析 Cuckoo monitor 中的 hook 脚本,获得 API 中 DLL 信息。本实验中共检测到 365 个 API、27 个 DLL。

本文获得了图 4 所示的 4 张元图,主要描述样本文件、API 及 DLL 的关系。关系 R_1 的矩阵 I 为 $3\,948 \times 365$ 大小的矩阵,行表示样本 F_i 对 365 个 API 的调用情况及调用的次数。根据 API 和 DLL 的对照,得到关系 R_2 的矩阵 B ,矩阵 B 为 365×27 维。

根据 2.2.1 节,按照图 4 所示的元图做有指导的随机游走。在实验过程中,为了获得更为丰富的信息,这样的随机游走需要进行多次才能刻画出一个样本点周围的节点信息,获得的 API 分

布才会尽可能接近样本的 API 实际分布,同时可以获得大量相关的样本文件作为辅助信息。

经过多次游走实验,最终确定 30 次随机游走可获得足够多的信息。因此,本文实验中规定任意的元图都将反复进行 30 次随机游走。以 S_1 为例,经过 30 次随机游走获得文件节点 F 周围的信息,即获得 30 组“F-A-F”信息。将这些信息统计到一个字典上,这个字典包含所有的 API、样本文件及 DLL,长度为 4 340 (即 $3\,948 + 365 + 27 = 4\,340$) 维。经过 30 次随机游走,可以获得 30 个字典长的向量,其语义信息是:按照规定的元图找到当前样本与周围节点的关系。以此作为样本根据这个子图获得的训练结果。

在 4 个元图上,都可以通过上述随机游走方法获得对应的上下文信息,最终会获得 4 个相应的词向量模型。

经过随机游走后,获得的词向量高达 4 340 维,为了达到降维的目的,本文采用 CBOW 模型对词向量降维,将 4 340 维字典嵌入 (embedding) 到 128 维中。

3.2 基于元图的分类结果分析

实验中,在每个元图指导下的随机游走获取的词向量经过降维后都可以独立用于样本分类。

1) 单元图分类测试

实验代码由 Python 语言编写,采用 kNN、RF (random forest) 及线性 SVM 分类器分别完成分类实验。所有实验均采用了十重交叉验证。

经多次实验验证,当 $k = 2$ 时,kNN 算法得到的分类结果最好,如表 1 所示。RF 分类器中决策树的数目设置为 10 时,结果如表 2 所示。SVM 分类器中 kernel 参数为“linear”(线性核)、惩罚参数 $C = 1$,分类结果如表 3 所示。

表 1 单元图模型 kNN 分类结果

Table 1 Classification results of each meta graph model using kNN				
评价指标	S_1	S_2	S_3	S_4
分类准确率	0.978	0.96	0.899	0.910
误报率	0.021	0.037	0.085	0.078
漏报率	0.022	0.041	0.101	0.090

表 2 单元图模型 RF 分类结果

Table 2 Classification results of each meta graph model using RF				
评价指标	S_1	S_2	S_3	S_4
分类准确率	0.942	0.923	0.869	0.858
误报率	0.053	0.067	0.104	0.112
漏报率	0.058	0.078	0.131	0.139

表 3 单元图模型线性 SVM 分类结果

Table 3 Classification results of each meta graph model using linear SVM				
评价指标	S_1	S_2	S_3	S_4
分类准确率	0.960	0.936	0.882	0.899
误报率	0.040	0.060	0.099	0.087
漏报率	0.040	0.063	0.118	0.100

表 1 ~ 表 3 中, S_1 、 S_2 、 S_3 及 S_4 分别为在图 4 中 4 个元图的指导下,根据改进的 Metagraph2vec 随机游走策略和 CBOW 模型获得的特征向量的分类结果。

可以发现,根据 S_1 元图游走获得的特征向量具有最好的分类准确率、最低的误报率和漏报率。这说明通过调用的 API 来判断样本之间的相似行为,具有非常好的鉴别能力。 S_2 元图中在寻找 2 个样本的关联关系中增加了 DLL 信息,也可以获得较好的分类结果。

S_4 元图相比 S_1 元图增加了通过 API 获取多样本之间关联性的游走。3 种分类器的分类结果表明,这种游走会对样本之间的相似性判断产生干扰。

根据 S_3 元图游走获取的特征向量具有最差的分类结果。 S_3 元图在 S_2 元图的基础上增加了 S_4 元图的游走路线,影响了分类效果。

2) 多元图融合的分类测试

为了综合应用这 4 个元图的分类结果,本文采用了投票的方法以便确定权重。在 4 个元图中分别对每个测试样本的词向量特征采用投票的方法确定样本属于哪一个家族,投票结果以百分比表示。通过主角度分析方法,确定各个元图分类结果的权重。

主角度 α 的计算方法为:假设有空间 Y_i 和 Y_j , 则其角度可以定义为

$$\cos \theta = \max_{\substack{\mathbf{y} \in Y_i \\ \mathbf{y}' \in Y_j}} \mathbf{y}^T \mathbf{y}' \tag{1}$$

式中: $\theta = 0$, 当且仅当 $Y_i \cap Y_j \neq \emptyset$; $\theta = \pi/2$, 当且仅当 $Y_i \perp Y_j$ 。

设 $\theta_1, \theta_2, \dots, \theta_d$ 为空间 Y_i 和 Y_j 主角度,则 Y_i 和 Y_j 的几何距离为

$$d(Y_i, Y_j) = \sqrt{\theta_1^2 + \theta_2^2 + \dots + \theta_d^2} \tag{2}$$

则权重 α_i 为

$$\alpha_i = \frac{\sum_{j=1, i \neq j}^m d(Y_i, Y_j)}{\sum_{i=1}^m \sum_{j=1, i \neq j}^m d(Y_i, Y_j)} \tag{3}$$

根据式(3)可以计算 4 个元图模型投票结果的权重,结果如表 4 所示。根据表 4 的权重,融合 4 个元图最终可获得如表 5 所示的分类结果。

表 4 元图权重结果				
Table 4 Weight values of meta graphs				
S	S ₁	S ₂	S ₃	S ₄
S ₁	0.000 16	1.659 9	0.586 8	1.820 5
S ₂	1.659 9	0.000 164 4	0.947 3	1.972 1
S ₃	0.586 8	0.947 3	0.000 231 4	0.905 5
S ₄	1.820 5	1.972 1	0.905 5	0.000 176 9
α	0.243 4	0.248 4	0.236	0.272 3

表 5 主角度融合分类结果			
Table 5 Classification results using principal angle hybrid method			
评价指标	kNN	RF	SVM
分类准确率	0.965 0	0.927 9	0.937 5
误报率	0.033 3	0.065 0	0.059 3
漏报率	0.034 4	0.072 3	0.063 5

表 5 中的分类准确率是 S_1 、 S_2 、 S_3 及 S_4 的融合结果(分类器的参数设置同单元图中的分类器参数设置)。相对表 1~表 3,表 5 的结果说明,融合的方法使得 S_3 及 S_4 元图的分类准确率得到了很大的提升。这说明,多元图相比单元图而言,可以对部分单元图的信息给与补充,可以通过 α 的物理意义得以解释。

3) 特征有效性分析

动态分析受运行条件、触发条件的影响较大,而恶意代码为了不被检测到,很多都增加了抗检测技术。因此,在 Cuckoo 沙箱中获取的 API、DLL 可能会存在漏报、误报的情况。为了分析 Cuckoo 沙箱获取的信息对本文方法的影响,采用 DynamoRIO 工具对 Cuckoo 沙箱中可执行恶意样本再次分析,重新获取 API、DLL 信息,并将二者结果做对比、合并。

在 3 个家族 3 948 个有效样本中 Cuckoo 沙箱共获得 365 个 API、27 个 DLL。DynamoRIO 再次分析这些样本,共获得 373 个 API、27 个 DLL 及 3 948 个有效样本。将 Cuckoo 与 DynamoRIO 结果合并之后,获得 3 948 个有效样本、375 个 API 及 28 个 DLL。

根据合并后的样本文件、API、DLL 信息构造异质网络,并再次完成单元图分类测试。表 6 为 $k=2$ 时 kNN 算法得到的分类结果。

表 6 合并 2 个沙箱结果后单元图模型 kNN 分类结果				
Table 6 Classification results of each meta graph model with two sandboxes results using kNN				
评价指标	S ₁	S ₂	S ₃	S ₄
分类准确率	0.978 3	0.957	0.895	0.911
误报率	0.021	0.036	0.083	0.078
漏报率	0.022	0.042	0.101	0.089

可以发现,表 6 与表 1 的结果基本一致。同样地,采用 RF 与 SVM 分类器可以得到与表 2、表 3 相似的实验结果,在此不再赘述。

从对比实验结果可以看到,本文采用的单元图指导下的多次随机游走策略,可以降低单特征(如 API、DLL)在异质信息网络中对分类结果的影响。

4) 与其他文献结果的对比

Fan 等^[11] 基于 Comodo 平台构建了 12 个元图,涉及到样本文件、API、DLL、压缩包、样本来源的机器(Machine)等信息,采用的是基本的随机游走策略及 skip-gram 方法描述样本特征。在该研究中,记录了单元图在 Comodo 公司提供的数据集上的分类准确率、召回率等,多元图融合后准确率可达 0.983;但是对于单元图(元路径)的准确率只有 0.75 左右。将 Fan 等^[11] 的方法应用到本文数据集上,在 S_1 元图的指导下,RF 的分类准确率只有 0.56,远远低于本文的单元图的分类准确率(见表 1)。这说明本文采用的随机游走和 CBOW 策略在单元图的分类效果是优于 Fan 等^[11] 的方法的。

由于 Comodo 公司的数据集并未公开,本文在 CNCERT 数据集上只能构造出有限种类的元图。在这种情况下,采取改进的 Metagraph2vec 随机游走策略与 CBOW 相结合的方式获得单元图特征向量描述,并取得了较好的分类准确率,这说明本文为信息类型不够丰富的异质信息网络的信息表示及提高单元图分类准确率方面提供了一个可行的研究方法。

4 结 论

本文从动态分析的角度分析恶意样本,将异质信息网络应用到恶意代码动态分析中,为恶意代码动态分析提供了一种新的特征构造、描述方法。本文实现了:

1) 通过在沙箱中运行样本,获得 PE 文件的动态访问信息,例如 API、DLL 等。通过构建异质信息网络的 4 个元图,获得样本文件与 API、样本文件与 DLL、API 与 DLL、样本文件与 API 及 DLL 之间的关系。

2) 提出改进的随机游走策略,并结合 CBOW 模型获得元图的特征描述。本文方法相比 Fan 等^[11] 的方法提高了单元图分类准确率。

3) 通过投票和主角度权重设置实现了多元图分类结果的融合。

实验中发现,多元图融合后的分类准确率提

高程度没有达到 Fan 等^[11] 给出的融合效果。经分析,这是由于本文异质信息网络节点类型不够丰富造成的,这将是下一步研究中要解决的问题。

参考文献 (References)

[1] SIKORSKI M, HONING A. Practical malware analysis: The hands on guide to dissecting malicious software[M]. San Francisco: No Starch Press,2012:1-2.

[2] 石川,孙怡舟,菲利普·俞. 异质信息网络的研究现状和未来发展[J]. 中国计算机学会通讯,2017,13(11):35-40.

SHI C,SUN Y Z,YU P. The research status and future development of heterogeneous information network[J]. Journal of China Computer Federation,2017,13(11):35-40(in Chinese).

[3] SHI C,LI Y,ZHANG J,et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering,2017,29(1):17-37.

[4] SUN Y Z,HAN J W,YAN X F,et al. PathSim: Meta path-based top-*k* similarity search in heterogeneous information networks [C]//Proceedings of the 37th International Conference on Very Large Data Bases,2011:992-1003.

[5] SUN Y Z,NORICK B,HAN J,et al. Integrating meta path selection with user-guided object clustering in heterogeneous information networks[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2012:1348-1356.

[6] SHI C,KONG X,HUANG Y,et al. HeteSim: A general framework for relevance measure in heterogeneous networks [J]. IEEE Transactions on Knowledge and Data Engineering,2014,26(10):2479-2492.

[7] CAO B,KONG X,YU P S. Collective prediction of multiple types of links in heterogeneous information networks[C]//Proceedings of the IEEE International Conference on Data Mining. Piscataway:IEEE Press,2015:50-59.

[8] HUANG Z,ZHENG Y,CHENG R,et al. Meta structure: Com-

puting relevance in large heterogeneous information networks [C] // Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2016:1595-1604.

[9] TAMERSOY A,ROUNDY K,CHAU D H. Guilt by association: Large scale malware detection by mining file-relation graphs [C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2014:1524-1533.

[10] CHEN L W,LI T,ABDULHAYOGLU M,et al. Intelligent malware detection based on file relation graphs[C] // Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing. Piscataway:IEEE Press,2015:85-92.

[11] FAN Y J,HOU S F,ZHANG Y M,et al. Gotcha-Sly malware! Scorpion: A Metagraph2vec based malware detection system [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2018:253-262.

[12] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2014:701-710.

[13] YU X D, CHAWLA N V, SWAMI A. Metapath2vec: Scalable representation learning for heterogeneous networks [C] // Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2017:135-144.

[14] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-09-07)[2020-09-01]. <https://arxiv.org/abs/1301.3781v3>.

[15] MIKOLOV T,SUTSKEVER I,CHEN K,et al. Distributed representations of words and phrases and their compositionality[C] // Proceedings of the 26th International Conference on Neural Information Processing Systems,2013,2:3111-3119.

Malicious code detection based on heterogeneous information network

LIU Yashu¹, HOU Yueran², YAN Hanbing^{3,*}

- (1. School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China;
2. Institute of Network Technology, Beijing University of Posts and Telecommunication, Beijing 100876, China;
3. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Malicious codes poses serious threats to network and information security. How to detect malware rapidly and how to eliminate and reduce the hazard caused by malware are important research topics. The paper presents a method to get dynamic features of malware using dynamic information and heterogeneous information network (HIN), and implements malicious codes detection and classification. Four meta graph schemes about FILE, API and DLL are proposed and malicious code HIN network pattern is described. An improved random walk strategy is used to obtain the context information of the object nodes in the meta graph schemes, which is used as the input of continuous bag of words (CBOW) model in order to get network embedding of word vectors. The method of principal angle is improved by voting to get the classification result of multiple meta graph schemes with feature fusion. The proposed method greatly improves the classification accuracy of malware based on the features of each meta graph when limited information is available.

Keywords: malicious code; heterogeneous information network (HIN); random walk; continuous bag of words (CBOW); meta graph

Received: 2020-09-23; Accepted: 2020-12-18; Published online: 2021-04-27 11:10

URL: kns.cnki.net/kcms/detail/11.2625.V.20210426.1819.003.html

Foundation items: National Key R & D Program of China (2018YFB0803604, 2018YFB0804704); National Natural Science Foundation of China (U1736218); the Fundamental Research Funds for Beijing University of Civil Engineering and Architecture (X20152)

* Corresponding author. E-mail: yhb@cert.org.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0414

加密文档排序中保序加密算法的最优化选取

张久岭*, 黄道超, 沈时军

(国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 对需要存储在服务器端的用户文档进行加密是对用户隐私保护的根本方法。采用不同的保序加密算法对加密文档排序结果有着不同的影响,寻找排序结果最优的保序加密算法是亟待解决的问题。提出了基于鉴别信息的保序加密算法选择的标准,比较了2种保序加密算法下对加密文档排序的性能。在通过保序加密算法得到的密文分布接近明文分布的情况下,排序得到的结果更接近明文检索的情况。所提出的选择标准具有理论上的意义,同时也可以指导在相同安全条件下保序加密算法的选择,以达到最优的检索结果。

关 键 词: 保序加密; Okapi BM25 模型; 鉴别信息; 加密信息检索; 用户隐私保护

中图分类号: TP399

文献标志码: A **文章编号:** 1001-5965(2022)02-0266-07

传统的电子邮箱服务、远程数据库服务等需要把包含用户个人隐私的数据存储在服务提供商的服务器上。云计算和云存储的快速发展降低了用户数据信息管理成本及存储资源与计算资源的消耗,而手持移动终端的愈加频繁使用使得更多用户数据存储在云端,这样在客观上造成了数据存储和计算资源的集中化。然而,随着越来越多的用户数据存储在远程服务器,对于服务提供商来说用户的数据是完全可见的。此外,服务提供商还可以对用户数据进行统计分析等操作,如果数据在不同的服务提供商间存在交互流动,则服务商可合并获取用户信息,会进一步侵犯用户隐私。因此,对用户隐私数据的保护越来越受到人们的关注,用户的数据隐私保护是一个迫切需要解决的问题。虽然可以通过签订服务协议或通过多方计算等方式避免服务提供商滥用用户数据,但是无法从根本上保护用户隐私。为有效地保护用户的隐私,对用户信息进行加密处理是一种解决方案。用户数据多以文档形式存储在服务提供商侧,因此对用户以文档形式进行信息加密并存储

在远程服务器是一种可行且有效的数据保护方案。

然而,当存储在远程服务器的用户加密数据形成规模之后,用户在进行查询以满足其信息需求时,需把加密文档全部下载回用户终端,这将带来巨大的通信开销,因此难以实现。如何对密文形式的数据进行检索成为亟待解决的问题。Song 等^[1]在 2000 年开创性地提出了一种实用的且可以搜索的对称加密算法,在进行加密关键词搜索时,需进行线性遍历,由于加密过程中一次一密,具有较强的抗统计分析能力。Boneh 等^[2]在 2004 年提出了一种可搜索公钥加密算法。Park 等^[3]提出了基于布隆过滤的多关键词安全搜索算法,布隆过滤在小规模数据集情况下可以进行有效检索,但在大规模数据集下,该算法检索时所需的布隆过滤器长度非常大。以上算法可以有效地从加密数据中检索到用户所需的文档。然而,随着数据集规模的增大,检索到可能相关的文档集合的规模也逐渐增大。例如,很多私人信息邮件被存储在某不可靠的邮件提供商服务器上,当进行检索时,即有对较多返回结果进行排序的需求。如

收稿日期: 2020-08-10; 录用日期: 2020-09-05; 网络出版时间: 2021-03-08 09:29
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210305.1810.001.html
基金项目: 国家重点研发计划(2016YFB1000105)
* 通信作者. E-mail: zhangjl@cert.org.cn

引用格式: 张久岭, 黄道超, 沈时军. 加密文档排序中保序加密算法的最优化选取[J]. 北京航空航天大学学报, 2022, 48(2): 266-272. ZHANG J L, HUANG D C, SHEN S J. Optimal choice of order preserving encryption scheme in encrypted document ranking [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 266-272 (in Chinese).

果将检索到的文档信息无差别返回给用户,一方面将带来巨大的通信、计算及存储开销,另一方面用户也无力消费大量的原始明文数据。因此,在不泄露用户信息的情况下,对检索到的加密文档进行相关性排序并给出最符合用户信息需求的文档是需要进一步研究和解决的问题。

在一般明文信息检索中,为对结果根据查询的相关性进行排序,一般利用到词频、关键词权重等关键信息。在加密信息检索情况下,为有效依据文档中的词频、关键词权重等加密信息对加密文档进行相关性排序,Swaminathan 等^[4]提出了基于保序加密算法的加密文档排序算法,该算法中文档的词频被保序加密算法加密,在需要进行排序时,把加密形式的词频相加给出相关度并以此进行排序。Wang 等^[5]提出了用保序加密算法对关键词权重进行加密,并对加密的权重进行加法运算得到查询与加密文档相似度的算法进行排序。Lu 等^[6]提出了将其类似的算法推广应用于多媒体检索中并对多媒体信息检索结果进行排序。Phong 等^[7]提出了基于保序加密算法的隐私保护深度学习系统,可以在不把个人数据透露给远程服务器的基础上,实现基于神经网络的深度学习。因此,有必要研究加密算法在加密文档排序中的应用。

在典型加密信息检索中,当查询中只有 1 个关键词时,通过基于保序加密算法的加密文档排序算法可以很好地对检索到的加密文档进行排序,这是因为用保序加密算法加密的明文数据加密后其密文顺序保持不变。然而,当查询中有 2 个及 2 个以上的关键词时,基于保序加密算法加密得到密文和并不一定保持与其明文和对应的顺序。在这种情况下,并非所有保序加密算法在多关键词查询时,其密文和均能较好保持顺序并取得较好排序结果。因此,可以通过适当地选择加密算法来保证密文和能最大限度保持其明文和的顺序,而研究探索最优化选取满足上述需求的加密算法也至少具有理论上的指导意义。

本文介绍了保序加密算法,拟合了 TREC 的标准数据集中关键词权重概率分布,比较了 2 种保序加密算法密文随明文的变化,提出了基于鉴别信息的保序加密算法选取标准。对选择不同保序加密算法的加密文档检索排序结果进行了比较,并在标准数据集上进行了实验,验证了采用开放式保序加密算法进行加密排序取得的结果优于采用封闭式保序加密算法的情况。

1 保序加密算法及其加密文档排序

1.1 保序加密算法

2003 年,Ozsoyoglu 等^[8]开创性提出了保序加密算法,以解决关系型数据库数值数据隐私保护的问题。2004 年,Agrawal 等^[9]提出了一种可以直接加密新数据且无需对已加密密文进行更新的保序加密算法。Boldyreva 等^[10]提出了基于超几何分布的保序对称加密算法及基于模的保序加密算法^[11]。文献[12]中的算法应用于无线传感网,通过加密信号数据并传输到传感节点,传感节点无需解密即可进行大小比较操作,降低了节点的功耗开销。在保序加密算法应用于远程数据库数据保护的相关研究中,有工作针对数据顺序本身也是敏感信息提出了隐藏密文顺序但支持有效比较运算的算法,提高了数据安全性^[13];也有工作提出了一种在数据库中插入假元组并保证数据完整性的算法,在该算法中真元组和假元组都经过保序加密,以允许用户验证保序加密数据库的查询完整性^[14];还有工作提出了具有有序选择重复明文分布攻击下的不可区分性的算法^[15]。其他算法包括基于一般近似公约数问题的保序加密算法^[16]及使用保序概率编码对明文的变换进行编码并通过 2 轮交互协议生成密文的公开保序密文生成算法^[17]。

由于上述提到的其他算法与开放式及封闭式保序加密算法具有本质上的相似性,为不失一般性,主要针对 Ozsoyoglu 等^[8]提出的 2 种典型保序加密算法进行分析,其第 1 种开放形式的保序加密算法实现过程的描述如下:

- 步骤 1 生成一个序列的随机数 $y_i \in [1, M]$ 。
- 步骤 2 定义 $z_i: S_1 := z_1 = y_1$ 。
- 步骤 3 对于 $k \geq 1$, 令 $A_k = M - S_k$ 。
- 步骤 4 $z_{k+1} = \text{Int}(y_{k+1} \times A_k / M)$ 。
- 步骤 5 $S_{k+1} = S_k + z_{k+1}$ 。

其中,对于明文 k ,保序加密算法加密的结果为 $f(k) = S_k$ (1)

第 2 种为封闭式的保序加密算法,其可以描述为

$$F(x) = C_n x^n + \cdots + C_1 x + C_0$$
 (2)

式中: x 为待加密变量; $C_i (0 \leq i \leq n)$ 为常量。

在其他算法中,Agrawal 等^[9]提出的算法把服从某一分布明文映射到服从另一分布的密文空间。而 Boldyreva 等^[10]提出的加密算法的实质是利用超几何分布产生保序的密文数据,与 Ozsoyo-

glu 等^[8]提出的第 1 种开放形式保序加密算法本质上类似。

1.2 明文文档检索及排序

典型的信息检索模型包括布尔模型、向量空间模型及概率模型等^[18]。布尔模型是基于布尔代数的检索模型,该模型中 1 个查询即构成了 1 个布尔表达式。尽管布尔模型在数据库检索中仍有使用价值,但在文档检索中无法对检索到的文档进行排序,在实践中使用较少。在向量空间模型中,文档及查询分别被标记为向量空间中的向量,通过向量之间的相似度来对文档进行相关性排序。概率模型通过计算文档与查询之间相关的概率对文档进行排序。为不失一般性,以概率模型的一种变形,即以 Okapi BM25 模型为例,实现对文档及查询相关度的计算。

Okapi BM25 模型自从 20 世纪 90 年代被提出后,由于检索的准确率相对其他模型较高,在文档排序中的应用非常普遍,并被推广应用到文本分类聚类等众多领域^[18-19]。

Okapi BM25 模型中,文档和查询之间的相似度定义为

$$\text{Sim}(Q,D)=\sum_{q_i\in Q}O(w_{q_i})\tag{3}$$

式中: w_{q_i} 为关键词 q_i 在文档 D 中所占的权重,定义为文档中词频和逆文档频率的函数,有很多个变种。本文中,为不失一般性,采用其中比较常用的一种:

$$w_{t_i}=\frac{3f(t_i,D)}{f(t_i,D)+1+\frac{|D|}{\text{avgdl}}}\lg\left(\frac{C-n(t_i)-0.5}{n(t_i)+0.5}\right)\tag{4}$$

式中: t_i 为关键词; $f(t_i,D)$ 为 t_i 在文档 D 中出现的频率; $|D|$ 为文档 D 的长度;avgdl 为文档集中平均文档长度; $n(t_i)$ 为文档集中出现关键词 t_i 的文档数; C 为文档集中文档的数量。

1.3 基于保序加密的文档排序

在加密文档检索中,当检索到所需的文档之后,需要对已经加密的文档进行基于词频或关键词权重对应密文的排序操作。Swaminathan 等^[4]提出了用保序加密算法对文档中的词频进行加密的方法,在查询单个关键词的情况下,通过保序加密的词频可以得到相关度的排序。尽管词频在一定程度上可以描述文档和查询的相关度,然而权重更为有效地描述了关键词在文档中的出现对文档与查询相关度的影响。

因此,通过对关键词的权重进行加密,可以更

为有效地描述一个关键词在文档中出现的作用。这里对文档排序的对象是关键词权重对应的密文。

理论上,当查询中关键词个数 $|Q|$ 为 1 时,用保序加密算法排序所得到的结果具有很好的性能。而当查询中关键词个数 $|Q|$ 多于 1 个时,由于涉及到对加密密文的加法操作,而加法后的密文排序并非明文和的顺序。因此,当查询中有多个关键词时,排序的结果与保序加密算法有关。

2 保序加密算法选取

2.1 明文分布

考虑到权重是词频及逆文档频率的乘积,在词频、逆文档频率是独立的简单假设情况下,那么二者的乘积是对数正态分布。然而,因为由词频决定的部分 $3f(t_i,D)/(f(t_i,D)+1+|D|/\text{avgdl})$ 与由逆文档频率决定的部分 $\lg((C-n(t_i)-0.5)/(n(t_i)+0.5))$ 并非是严格独立的,所以难以得出 w_{t_i} 分布的封闭表达式。通过在 TREC 的标准数据集 ClueWeb09 Cat. B 上对各个权重进行统计(ClueWeb09 是国际信息检索领域研究使用的标准化数据集),给出了其实际分布,并依据类似对数正态分布密度的函数对其分布进行了拟合。ClueWeb09 Cat. B 中的关键词权重分布如图 1 所示。此处,每个关键词权重的值均分布于 $[0,3)$ 。由于分布由文档集的很多因素共同决定,对其进行拟合的结果为

$$f_{\text{lognormal}}(x)=p_1x^{p_2}\exp[-p_3(\ln x-p_4)^{p_5}]\tag{5}$$

式中:各参数的拟合值如表 1 所示。

由图 1 可以看出,该分布并非严格的对数正态分布,这是因为词频和逆文档频率并非是完全独立的,而词频本身是对关键词出现频率进行运算的结果,且权重的分布是选择保序加密算法的决定性因素。

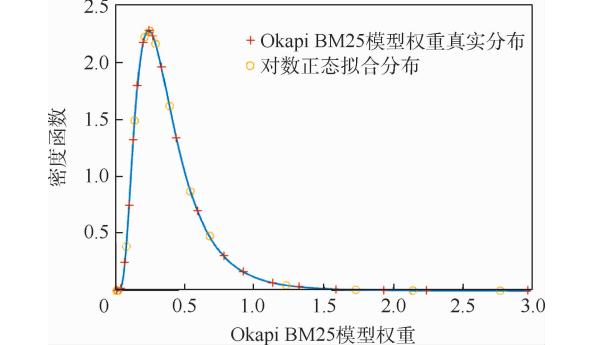


图 1 Okapi BM25 模型权重出现频率的分布函数
Fig. 1 Probability distribution function of term weights in the Okapi BM25 model

表1 参数拟合值
Table 1 Parameter fitting values

参数	拟合值	95% 置信区间
p_1	0.500	(0.495, 0.506)
p_2	-1.356	(-1.368, -1.345)
p_3	1.448	(1.442, 1.455)
p_4	-0.960	(-0.964, -0.957)
p_5	1.419	(1.414, 1.424)

2.2 保序加密算法选取方法

对2种保序加密算法下的密文进行了比较,结果如图2所示。由于加密后的值较大不便于直接展示,为便于比较,对密文取对数。图2的曲线表明,当采用开放形式保序加密算法的密文值增大到一定程度后便保持稳定,而封闭形式保序加密算法的密文值一直明显地增加。

2种保序加密算法都可以很好地对加密数值保持顺序。然而,本文有以下启发性结论:针对密文的加法运算,采用开放形式保序加密算法加密的密文能更好地保持其顺序。这是因为:用开放式保序算法加密得到的密文值比较接近,然而,用封闭式加密算法加密得到的密文值在数值空间上分布越来越稀疏。因此,封闭式保序加密算法加密得到的密文值加法运算结果的保序效果劣于开放式保序加密算法。反映在加密文档排序效果上,采用开放形式保序加密算法的检索结果在性能上优于封闭式保序加密算法得到的检索结果。

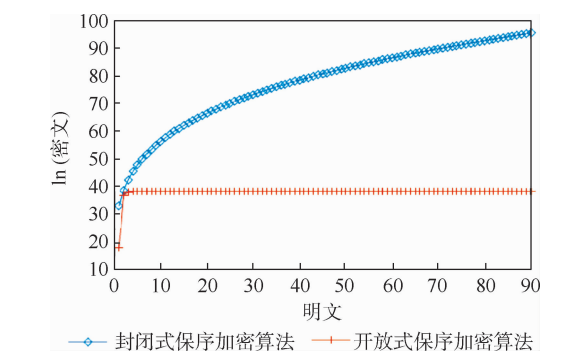


图2 开放式保序加密算法与封闭式保序加密算法密文随明文变化情况

Fig.2 Change of cipher text with plaintext for open order preserving encryption scheme and closed order preserving encryption scheme

3 保序加密算法选取标准

理想情况下,设保序加密由连续函数 $h(x)$ 表示,那么,给定明文 X ,则其对应密文 $Y = h(X)$ 的分布为

$$g(y) = \frac{f(h^{-1}(y))}{h'(h^{-1}(y))} \tag{6}$$

为使在保序加密的情况下其加法在统计意义上仍然是保序的,通过保序加密算法加密的密文如果服从明文分布 $f(x)$ 或接近明文的分布,则在理论上同样可以达到明文检索下相差较小的检索结果。经过加密变换 $Y = h(X)$ 之后, Y 值的变化区间为 $[h(0), h(3))$,为便于进行分布比较,可以将 Y 通过某线性变换 $Z = 3(Y - h(0))/(h(3) - h(0))$ 将其值重新映射到区间 $[0, 3)$ 。线性变换不影响2个数加法的结果,因此,可以对 Z 与 X 的分布进行比较。 Z 的分布由 Y 的分布表示为

$$p(z) = \frac{h(3) - h(0)}{3} g\left(\frac{h(3) - h(0)}{3} z + h(0)\right) \tag{7}$$

衡量2种分布之间差异性的一种比较好的方法是鉴别信息。通过鉴别信息可以对明文与密文概率分布的差异性进行描述。为使二者分布距离最近,需要使鉴别信息最小。为使保序加密后的密文仍然保持原有的分布,需要满足:

$$\min_{h(\cdot)} \left\{ \int_0^3 f(x) \lg \frac{f(x)}{p(x)} dx \right\} \tag{8}$$

然而,对保序加密算法的选择是有其他约束性的。例如,为保证算法的安全性,封闭式保序加密算法系数的选取范围应该足够大;而在开放性的保序加密算法中,为保证明文对应的密文各不相同,需要增加密文值域等。因此,保序加密算法 $h(x)$ 的优化选择可为

$$\begin{aligned} \hat{h}(x) &= \arg \min_{h(x)} \left\{ \int_0^3 f(x) \lg \frac{f(x)}{p(x)} dx \right\} \\ p(x) &= \frac{h(3) - h(0)}{3} g\left(\frac{h(3) - h(0)}{3} x + h(0)\right); \\ g(x) &= \frac{f(h^{-1}(x))}{h'(h^{-1}(x))}; h(x) \in H \end{aligned} \tag{9}$$

式中: H 为满足一定安全性能的保序加密算法的集合。式(9)可以作为保序加密算法选取的准则。

在实际情况中,由于保序加密算法受一系列因素的制约,难以由式(9)给出关于 $\hat{h}(x)$ 的封闭表达式。此外,加密算法 $Y = h(X)$ 在一般情况下其反函数也难以求解,导致对 $\hat{h}(x)$ 的求解更加困难。尽管计算 $\hat{h}(x)$ 存在着困难,式(9)仍然有理论上的指导意义。而且,在给定2个或多个保序加密算法的情况下,为最优化选择提供了标准。

4 实验结果

有工作分析研究了云数据搜索中一到多保序加密算法的安全性,并提出了通过保序密文的差异可以进行差分攻击的问题^[20]。本文从另一个

角度,即在算法具有同等安全性的前提下对如何最优化选取保序加密算法进行了研究。在 TREC 的标准数据集 ClueWeb09 Cat. B 上进行了实验,由于数据集规模较大,其结果相比 Cranfield 等小规模数据集更接近真实环境。在该数据集中,有 50 000 000 个文档,50 个标准查询。针对该数据集的任 1 个文档进行分词、记干化等预处理,可以得到其对应的词频 $f(t_i, D)$ 及文档长度 $|D|$ 。然而,通过对整个文档进行统计,可以得到式(4)中参数的 $avgdl$ 、 C 及 $n(t_i)$ 的值。

通过以上参数,可以计算得到关键词 t_i 所对应的 Okapi BM25 模型下的权重。用户对其通过保序加密算法进行加密,把得到的密文存储在服务器端,而在服务器端,对对应关键词的保序加密形式的权重进行加法运算,通过所得到的结果对文档进行排序。

在实验中所采用的保序加密算法分别为 1.1 节中文献[8]所提到的开放式保序加密算法与封闭式保序加密算法。

评价检索结果效果的指标为准确率及返回率。准确率定义为返回的且相关的文档占所返回文档的比例,返回率定义为返回的且相关的文档占有所有相关文档的比例。

采用不同保序加密算法对加密文档进行排序及明文检索情况下的准确率-返回率曲线的对比结果如图 3 所示。可以看出,在 ClueWeb09 Cat. B 数据集上,采用保序加密算法排序的准确率-返回率曲线比明文检索的情况低;采用开放式保序加密的准确率-返回率曲线比封闭式保序加密对应的曲线高。

前 N 个文档的准确率($P@N$)如表 2 所示,此处 N 的取值分别为 5、10、20、30、50、100。表 2 中不同返回文档个数下的准确率表明,开放式的保序加密算法可以达到较封闭式保序加密算法更

表 2 前 N 个返回文档的准确率

Table 2 Precision of top N recalled documents

$P@N$	准确率/%		
	开放式保序加密算法	封闭式保序加密算法	明文检索
$P@5$	63.6	57.2	78.0
$P@10$	54.0	50.6	74.6
$P@20$	46.8	42.0	67.6
$P@30$	42.9	38.7	62.1
$P@50$	38.0	35.0	56.2
$P@100$	30.5	29.2	43.5

好的检索结果,比较重要的指标,如前 10 个及前 20 个的准确率明显高于后者,其中 $P@10$ 高出 3.4%。然而,2 种保序加密算法的准确率都低于作为比较基准的明文检索情况下的检索排序准确率。这说明单纯依靠保序加密算法可以开展对加密文档的排序,但距离明文检索下的结果还有差距,后续将研究通过结合使用同态加密算法来进一步提高检索的准确率。通过结合使用保序加密及同态加密算法,可在减少在本地运算量的基础上,为密文空间提供更加精细的相关度计算结果。

5 结 论

1) 通过在 ClueWeb09 Cat. B 数据集上对权重进行统计分析,拟合了其权重的分布。

2) $P@10$ 的加密文档排序中,基于开放式保序加密算法的准确率比封闭式保序加密算法高出 3.4%,验证了在加法运算下的开放形式保序加密算法的密文能更好地保持原有顺序的结论。

3) 给出了对保序加密算法最优化选取的准则,该准则具有理论上的指导意义。

探索保序加密算法的共性特征及在此基础上提出最优化选取方案有待进一步的深入研究。

参考文献 (References)

[1] SONG D X, WAGNER D, PERRIG A. Practical techniques for searches on encrypted data[C] // Proceedings of 2000 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2000:44-55.

[2] BONEH D, CRESCENZO G, OSTROVSKY R, et al. Public key encryption with keyword search[C] // Proceedings of Advances in Cryptology-Eurocrypt 2004. Berlin: Springer, 2004:506-522.

[3] PARK D J, KIM K, LEE P J. Public key encryption with conjunctive field keyword search[C] // International Conference on Information Security Applications. Berlin: Springer, 2005:73-86.

[4] SWAMINATHAN A, MAO Y N, SU G M, et al. Confidentiality preserving rank-ordered search[C] // Proceedings of the ACM Workshop on Storage Security and Survivability. New York:

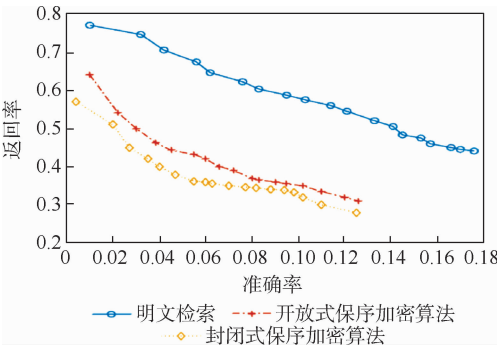


图 3 不同检索排序方式下的准确率-返回率曲线

Fig.3 Curves of precision versus recall rate with different retrieval and ranking methods

ACM,2007:7-12.

[5] WANG C,CAO N,LI J,et al. Secure ranked keyword search over encrypted cloud data[C] // Proceedings of 30th IEEE International Conference on Distributed Computing Systems. Piscataway:IEEE Press,2010:253-262.

[6] LU W J,SWAMINATHAN A,VARNA A L,et al. Enabling search over encrypted multimedia databases[C] // Media Forensics and Security. Bellingham:SPIE Press,2009:725418.

[7] PHONG L T,AONO Y,HAYASHI T,et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security,2018,13(5):1333-1345.

[8] OZSOYOGLU G,SINGER D,CHUNG S. Anti-tamper databases:Querying encrypted databases[C] // Proceedings of the 17th Annual IFIP WG. New York:ACM,2003:4-6.

[9] AGRAWAL R,KIERNAN J,SRIKANT R,et al. Order preserving encryption for numeric data[C] // Proceedings of the ACM SIGMOD International Conference on Management of Data. New York:ACM,2004:563-574.

[10] BOLDYREVA A,CHENETTE N,LEE Y,et al. Order-preserving symmetric encryption[C] // Proceedings of Advances in Cryptology-Eurocrypt 2009. Berlin:Springer,2009:224-241.

[11] BOLDYREVA A,CHENETTE N,O' NEILL A. Order-preserving encryption revisited:Improved security analysis and alternative solutions[C] // Advances in Cryptology-CRYPTO 2011. Berlin:Springer,2011:578-595.

[12] KHOURY E,MEDLEJ M,JAOUDE C A,et al. Novel order preserving encryption scheme for wireless sensor networks[C] // IEEE Middle East and North Africa Communications Conference (MENACOMM). Piscataway:IEEE Press,2018:1-6.

[13] AHMED S,ANNIS A,ZAMAN A,et al. Semi-order preserving encryption technique for numeric database [J]. International Journal of Networking and Communication,2019,9(1):111-129.

[14] SHEN N,YEH J,CHEN C,et al. Ensuring query completeness in outsourced database using order-preserving encryption[C] // Proceedings of IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking. Piscataway: IEEE Press, 2019: 776-783.

[15] REDDY K S, RAMACHANDRAM S. A secure, fast insert and efficient search order preserving encryption scheme for outsourced databases[J]. International Journal of Advanced Intelligence Paradigms,2019,13(1/2):155-177.

[16] DYER J,DYER M,DJEMAME K. Order-preserving encryption using approximate common divisors[J]. Journal of Information and Applications,2019,49:102391.

[17] CHOWDHURY A R, RAMANATHAN P. Public order-preserving cipher generation scheme for distributed computing[C] // Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York:ACM,2018:2273-2275.

[18] MANNING C D,RAGHAVAN P,SCHUTZE H. Introduction to information retrieval [M]. Cambridge: Cambridge University Press,2008:253-256.

[19] WHISSELL J S,CLARKE C. Improving document clustering using Okapi BM25 feature weighting[J]. Information Retrieval, 2011,14(5):466-487.

[20] LI K,ZHANG W M,YANG C,et al. Security analysis on one-to-many order preserving encryption-based cloud data search [J]. IEEE Transactions on Information Forensics and Security, 2015,10(9):1918-1926.

Optimal choice of order preserving encryption scheme in encrypted document ranking

ZHANG Jiuling^{*}, HUANG Daochao, SHEN Shijun

(National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Encrypting the documents before uploading them to the untrusted removable server is one of the ultimate solutions to protect the user privacy. As different encryption schemes give different ranking results, finding the order preserving encryption scheme that gives the best ranking result is the key issue to be solved. A discrimination information based on criterion for selecting the order preserving encryption scheme is proposed and then employed to compare the ranking results of two different order preserving encryption schemes. In the case the distribution of the cipher text is closer to that of the plaintext, the ranking result is closer to that in the plaintext scenario. The proposed criterion is not only instructive in theory, but also helpful practically in choosing the order preserving encryption scheme under the same safety conditions, which further lead to the optimal retrieval result.

Keywords: order preserving encryption; Okapi BM25 model; discrimination information; encrypted information retrieval; user privacy protection

Received: 2020-08-10; **Accepted:** 2020-09-05; **Published online:** 2021-03-08 09:29
URL: kns.cnki.net/kcms/detail/11.2625.V.20210305.1810.001.html
Foundation item: National Key R & D Program of China (2016YFB1000105)
*** Corresponding author.** E-mail: zhangjl@cert.org.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0388

电信大数据分析下的时空区域经济可视化应用

李娜, 刘文敏*, 孟繁瑞, 刘岩

(国家计算机网络应急技术处理协调中心山东分中心, 济南 250002)

摘 要: 当前, 国内移动电话用户已达 15.9 亿, 在巨大的用户基数下, 电信大数据呈现的特征在一定程度上反映了人群活动特征, 进一步能够反映特定区域的发展状况。时空区域经济可视化应用利用数据挖掘技术对电信大数据进行处理和提取, 以提高数据质量, 并对数据进行不同规则的筛选, 通过建模技术进行分析, 结合电子地图数据、交通数据等多源信息, 多角度分析用户行为特征。该应用分析对时空区域经济状况进行可视化研究, 分析居民生活属性, 同时, 利用双重差分(DID)统计模型对区域经济政策进行评价。基于特征分析结果, 为区域经济发展热点选址、指导城市商圈布局提供决策依据, 提高了城市系统运行的效率, 扩大了经济区域效益范围。

关 键 词: 电信大数据; 手机信令; 数据挖掘; 区域经济; 可视化

中图分类号: TP391

文献标志码: A **文章编号:** 1001-5965(2022)02-0273-09

随着移动互联网、智能手机、大数据等新一代技术的发展, 信息化技术日益普及, 极大程度上改变了居民的生活方式。社会信息化和网络化发展导致数据量激增, 数据的规律性特征在一定程度上反映了人群在现实生活中的活动特征。在“以人为本”的社会中, 人群的行为特征和方式也能够反映出特定区域的发展状况^[1-2]。大数据技术时代, 国内越来越多的城市提出要加强“智慧城市”建设, “智慧城市”通过有效整合多源信息资源, 为城市规划、建设和管理构建新的模式。“智慧城市”建设过程中, 如何处理、整合海量数据并加以利用成为了关键。同时, 在区域发展规划及经济趋势预测中, 大数据的运用已成为新的研究方式, 该方式有别于过去普遍采用的问卷调查方式。传统区域经济的研究数据来源多通过人口普查和经济调查的方式进行, 其实时性低、泛化性低、准确性低等缺点难以避免; 在数据层面, 其体量小、不全面、质量参差不齐, 获取数据的人力成

本高、区域覆盖范围小, 且多为政府人员采集, 非公开性的特点更是为经济预测带来重重困难。而大数据的 5V 特征, 即规模性 (volume)、多样性 (variety)、高速性 (velocity)、价值性 (value) 和真实性 (veracity), 一定程度上克服了传统方式的缺点, 其数据量大、获取统一、成本低等优点极大地弥补了传统数据的不足。在众多可用的海量数据中, 基于通信网络的电信大数据更具有得天独厚的优势, 其时空全覆盖的特点是其他数据所不拥有的。数据的时间连续性能够在历史数据和当前数据的基础上对未来数据进行预测与趋势分析; 空间多样性能够在广阔的覆盖面上, 动态、实时、快速地进行更新、共享、融合分析。《中国移动互联网发展报告(2020)》指出, 截至 2019 年 12 月底, 中国 4G 手机用户数量达到 12.8 亿, 占移动电话用户总数的 80.1%^[3]。利用数据挖掘技术, 能够刻画覆盖用户人群的行为画像, 提取行为特征^[4-5], 结合区域经济政策进行评价, 可为经济发

收稿日期: 2020-08-04; 录用日期: 2020-09-05; 网络出版时间: 2020-11-17 15:35
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201117.1416.001.html
* 通信作者。E-mail: wenminabe@qq.com

引用格式: 李娜, 刘文敏, 孟繁瑞, 等. 电信大数据分析下的时空区域经济可视化应用[J]. 北京航空航天大学学报, 2022, 48(2): 273-281. LI N, LIU W M, MENG F R, et al. Application of space time regional economy visualization based on telecom big data analysis[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 273-281 (in Chinese).

展提供决策依据。

本文采用多源数据融合分析的方法,在充分研究遥感影像技术^[6-7]、手机信令技术^[8-9]、交通热力图技术^[10]等的基础上,将传统数据与电信大数据融合,实现在电信大数据下的区域经济分析,本文主要工作如下:

- 1) 通过分析数据特征,利用数据挖掘技术完成数据清洗和时空关系关联,获取并补充电信大数据中精确的时间戳、地址等属性,提高数据质量。
- 2) 结合多源数据,对数据特征规则进行筛选和建模,定量分析关联特征,设计基于电信大数据的区域经济分析方法框架,全方位、多角度地分析区域经济状况。
- 3) 对电信大数据下的时空区域经济进行可视化实验分析,结合优化选址方法,指导城市热点商圈布局与调整,挖掘投资商业价值,提高城市系统效率,为“智慧城市”规划提供参考。

1 电信大数据清洗分析

本节对电信大数据进行清洗分析,并选取其中关键项进行数据设计和分析处理。

电信大数据数据量庞大,以 2019 年 9 月某省 1 日内产生的数据为例,数据总量达 6TB,涵盖约 9 000 万用户产生的 700 亿条记录。其数据格式如表 1 所示,部分关键项包含用户标识、基站标识码、位置信息、时间戳、手机归属地、手机所在地、事件标识等信息。

对海量的电信大数据进行清洗、去噪和纠偏尤为重要。研究过程中,利用 Spark 集群计算框架与 HDFS 文件存储框架,对大量原始的数据进行了预处理,采用 Hive 存储各类中间结果,采用 Hbase 存储用于可视化的大规模数据,采用 Redis 快速存取各类配置和用于界面实时展示数据量较少的分析结果,进而完成了数据的过滤清洗。

表 1 电信大数据基本格式

Table 1 Basic format of telecom big data		
序号	字段名称	字段描述
1	USER_ID	用户标识
2	ULI	基站标识码
3	LAC	位置区域编号
4	CI	位置小区编号
5	APPEAR_TIME	时间戳
6	DURATION	停留时间
7	HOMECODE	用户归属地
8	AREACODE	手机所在地
9	DATE	时间日期
10	HOURL	小时
11	EVENT_FLAG	事件标识

1.1 无效数据清洗

原始数据中存在部分无效数据,包括空值、重复值、错误值及格式错误数据等,占比约为 0.8%。对于存在极小概率的某些属性字段中会出现缺失信息,采用直接删除的处理方法;对于某些属性字段完全一致的重复数据,保留其中一条数据,删除重复数据,进一步减小计算量;对于存在偏差范围无法覆盖的异常数据及格式错误数据(如不在指定区域经纬度范围内的数据、日期异常、停留时间异常等),采用条件筛选或聚类算法进行清洗。

1.2 噪点数据清洗

原始数据中的人群位置信息是用户处于手机联网时连接的基站位置,由于终端和基站之间通信复杂、时有干扰,在基站比较密集的地区其手机上报的基站附着点会有很多噪声,对基站附着点进行去噪,可以提升数据质量,有利于后续数据分析。

首先对数据进行去重操作,按取 10 位 Geo-Hash 算法^[11]对数据点合并,利用空间密度聚类 DBSCAN 算法^[12-13]对附着点区域直径 50 m、样本量少于 5 的孤点去噪,利用 K-Means 算法^[14]进行聚类,按聚类类别进行排序,将同类数据合并,得到去噪后的数据结果。同一基站 1 d 内的数据处理前,基站附着点有 4 430 个,如图 1 所示。对其进行数据点去噪,共去除 820 个,保留 3 610 个,保留 81.49% 数据体量,去噪后基站附着点如图 2 所示。去除冗杂的数据点后,提升了数据质量。

1.3 轨迹纠偏

在基站比较密集的地区,手机附着的基站其连接轨迹点会来回跳转,使局部轨迹呈现网状结构,并产生乒乓数据^[15-16],轨迹纠偏变得更加复杂。对基站轨迹的纠偏去噪,可以从 3 个方面进行处理。



图 1 数据清洗前的基站附着点分布

Fig.1 Distribution of base station attachment points before data cleaning

依次进行 7 位的 GeoHash 算法^[11]取坐标值,对异常数据纠偏、速度异常数据纠偏、乒乓数据降噪纠偏。舍弃提取的误差范围以外的异常数据,得到清洗后的轨迹数据。轨迹纠偏流程如图 3 所示。

图 4 和图 5 为纠偏前后对比,该轨迹为放置在公交车上的用户手机在相同线路行驶 2 个来回、共 4 趟距离的轨迹路程。其原始轨迹如图 4 所示,轨迹点有 879 个,对其进行轨迹纠偏,共去除 94 个,保留 785 个,保留 89.31% 数据体量,纠偏后轨迹如图 5 所示。对比图直观地显示出,移除明显的漂移轨迹点可以提高数据的聚敛性,去除乒乓数据影响后的网状结构,轨迹更加平滑,为后面的分析提供更准确的数据支撑。

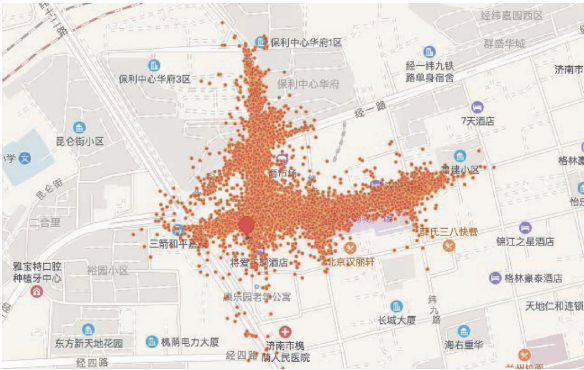


图 2 数据清洗后的基站附着点分布

Fig. 2 Distribution of base station attachment points after data cleaning

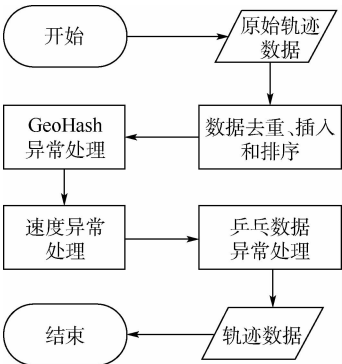


图 3 轨迹纠偏流程

Fig. 3 Flowchart of track rectification



图 4 数据纠偏前的轨迹

Fig. 4 Trajectory before data correction



图 5 数据纠偏后的轨迹

Fig. 5 Trajectory after data correction

2 基于电信大数据的区域经济分析方法

区域经济在一定程度上反映了居民的生活水平。基于电信大数据进行规则筛选并对其特征建模,结合电子地图、交通出行数据、人口普查数据等多源数据,进行多尺度、多维度分析,可以为区域经济决策提供参考,具体方法框架如图 6 所示。在大量原始数据基础上,通过数据清洗提高数据质量,在定义规则中选取特定时间和区域,精确缩小时空范围,进行时空关联并分析。通过数据结果将其可视化,利用图表模式展示分析,将数据通过计量经济学模型进行政策评价,对分析居民社会生活属性、经济发展、政策区域选址等工作提供了有力的数据支撑。本节选取特定时空区域,以某市“夜经济”为例进行区域经济分析方法框架设计。

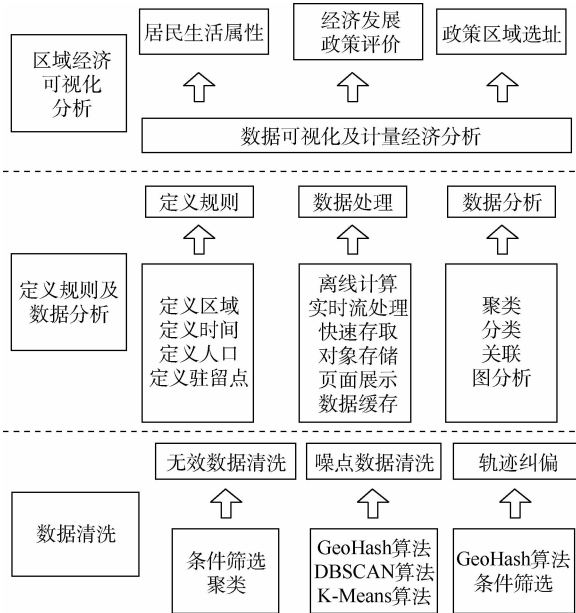


图 6 电信大数据的区域经济分析方法框架

Fig. 6 Framework of regional economic analysis method for telecom big data

夜经济起源于英国,1995 年被英国政府纳入城市发展目标^[17],2017 年,夜经济占英国 GDP 的 8%,涵盖约 1/8 的就业岗位。在中国,城市夜经济经历了 3 个发展阶段,即初始市场、零散夜市和集成商圈^[18]。许多一二线城市在 2019 年纷纷制定夜经济发展战略,并有向三四线城市扩散的迹象。某市在 2019 年 6 月出台了《关于推进夜间经济发展的实施意见》,同时进行了大规模的灯光亮化工程、定点夜市鼓励政策等,使夜经济快速升温。2019 年 11 月,首届中国夜间经济论坛上,该市入选“夜间经济十佳城市”。2020 年 5 月,中国受疫情影响,为恢复消费活力,国家再次开放发展夜市经济,推出了“五允许一坚持”政策,下文将基于电信大数据并结合相关政策进行有效评价。

首先,针对某市“夜经济”定义规则,结合已有电信大数据、用户轨迹、人物画像库等信息,提取数据特定属性。其次,划定空间区域范围,选取特定时间范围,查找范围内对应用户轨迹、结合人物画像知识库属性建模,通过 Hadoop 平台进行分析。最后,对夜间活跃人口行为、出行游玩地点、交通情况等连同经济政策进行评价,并提出选址优化方法,指导城市热点商圈布局与调整,有助于政府决策区域优化选址方法。

2.1 数据规则定义

- 1) 夜间经济:18:00 至次日 6:00 城市各种商业经营活动的总称。
- 2) 夜间人口:18:00 至次日 6:00 活动的人口。
- 3) 夜间活跃人口:夜间发生 2 km 以上且排除前往火车站位置的移动人口。
- 4) 夜间人口驻留点:夜间出现停留时长在 0.5 h 以上及 5 h 以下的位置点。
- 5) 某景点夜间游玩人数:某日夜间在景点基站附近,提取停留时长大于 0.5 h 且为外地归属信号与停留时长大于 0.5 h 且为居住地不在该景点的信号,二者加和。
- 6) 夜间人流走廊:夜间活跃人口在某市区迁移情况。
- 7) 外出情况:每小时与常居住地距离在 2 km 以上的连接数数量统计。
- 8) 夜间人口聚集点识别:提取夜间某时间段内终端所在地,组点形成分布,排除与家庭住址一致的终端,进行聚类。
- 9) 区域范围:某市。

通过对以上规则可视化分析,通过图、表、热力图等对该市区域夜经济情况进行剖析,具体见

第 3 节电信大数据下时空区域经济可视化实验分析。

2.2 电信大数据下时空区域经济发展政策评价算法

地方政策实施效果评估是微观计量经济学的前沿主题。2019 年 6 月,中国出台了《关于推进夜间经济发展的实施意见》,针对夜间的外出意愿和夜经济政策的相关性进行统计学分析,采用双重差分(difference-in-differences, DID)模型算法^[19],通过控制实施组和对照组在政策处理前后存在的差异实行政策评价。

DID 是评估一项政策是否有效的重要算法,特点为事前差异,即除了要研究的核心变量差异外,还包含产生干扰的其他因素。DID 模型的基础是自然实验数据,即被解释变量是由带参数的连续解释变量、离散解释变量控制的。通过建模,控制事前差异,得到真实结果:

$$Y = \alpha D_i + \beta X + \gamma (D_i X) + u \tag{1}$$

式中:Y 为被解释变量;X 为连续解释变量;D_i 为离散解释变量;u 为截距项,截距项表示解释变量为 0 时被解释变量的值,在回归算法中意义不大。

式(1)中重点关注 α、β、γ 的斜率变化,斜率的正负反映出解释变量和被解释变量的正负相关性,斜率为正则代表两变量正相关,斜率为负则代表两变量负相关。

2.3 政策区域选址

将特定时间内的经济活跃区域景点人流数、人口出游数、交通拥堵区域位置、电信大数据呈现的人流密度等数据整合分析,划定时空区域关联分析,利用专家判断、趋势分析、模拟等方法进行经济活跃区域和热门地点输出预测,如特定法定节假日、每日特定时段、重要时间节点等相关区域分析。有助于政府对相关时段的区域经济发展做出决策,选取合适的热门地点有助于市民对市场经济响应。

3 电信大数据下时空区域经济可视化实验分析

3.1 实验环境与数据

3.1.1 实验环境

对海量原始的数据进行处理并建模,分析和计算建模结果。应用技术工具主要为 Spark、Spark streaming、Hive、Hbase、Redis、Kafka 等,其中 Spark 用于运行离线计算程序,Spark streaming 用于进行实时的流处理,Hive 用于存储各类中间结

果,Hbase 用于存储可用于界面展示的大规模数据库,Redis 用于快速存取各类配置和用于界面实时展示的数据量较少的分析结果,Kafka 用于在各个程序模块之间实时发布和订阅消息。用于实验的服务器及其配置如表 2 所示。

表 2 实验环境
Table 2 Experimental environment

服务器	配置
系统	CentOS 7.3
内存	256 GB
磁盘	7.3TB×6
CPU	Intel(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz
开发语言	Java、Python、Scala 等
系统技术框架	Spark 集群计算框架、HDFS 文件存储框架、Hive、Hbase、Redis、Kafka 等

3.1.2 实验数据

- 1) 数据来源。电信大数据及部分政府统计年鉴数据。
- 2) 数据周期。主要数据来源于 2019 年 9 月至 10 月,部分数据来源于 2019 年 12 月。
- 3) 数据处理。利用集群计算框架及文件存储框架,对原始数据进行去重清洗处理,将原始数据解释为区域化数据、轨迹数据和停留数据,并进一步分析。获取轨迹停留数据后,对数据进行基站圈点,得到局部数据,按照时间点、地点等需求打点,利用 pyecharts 等相应的第三方工具包,在地图上根据经纬度和量值,离线画出热力图等,呈现可视化分析结果。

3.2 时空区域经济可视化分析

基于提取数据特征信息,结合交通、天气等多源数据,将可视化分析形成报告,分为 5 个方面:夜间活动总概况、游在夜间、行在夜间、玩在夜间及政策评价和结论。

3.2.1 夜间活动总概况

- 1) 夜间活跃人口。随着夜晚来临,居民活动总体趋势下降,22:00 左右出现骤降,23:00 有 32 万人在外活动,如图 7 所示。将 2:00—3:00 点活动人群视作夜间上班人群,进行差分计算,22:00 仍有超 15 万人在外休闲娱乐。活跃人口随时间分布符合事实规律,进一步验证了数据的准确性。
- 2) 夜间活跃人口年龄和性别比例。夜间活跃在外人群以男性为主(67.18%),年龄分布集中在 15~40 岁,如图 8 所示。一方面是年轻人相对有活力,另一方面是男性比女性拥有更强的夜间外出休闲、娱乐意愿。
- 3) 不同行政区活跃人口及人口流动情况。热闹的夜晚正改变着夜生活,从某市范围看,居住

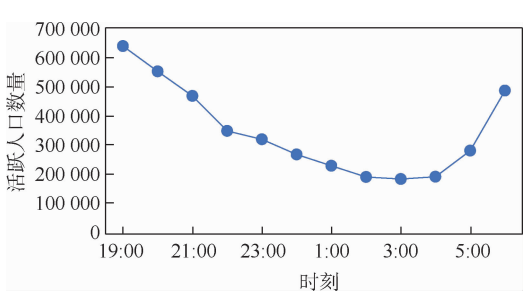


图 7 九月夜间活跃人口
Fig. 7 Active population at night in September

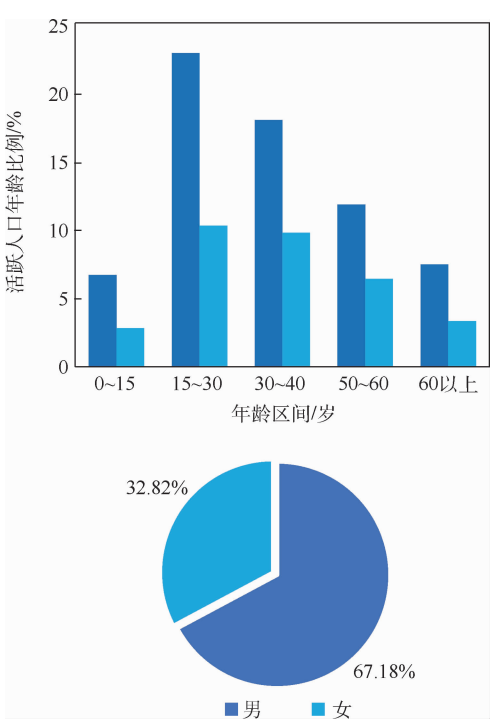


图 8 夜间活跃人口年龄和性别比例
Fig. 8 Age and gender ratio of active population at night

在历城区居民活跃人口最多,而从居住地与活跃人口比例上来说,某市槐荫区、历下区和天桥区占比前 3,最不活跃的是章丘区,图 9 反映了某市各区的夜间经济活跃情况。

从夜间人口迁移数量中可以看出,某市历城区是整个人口流动网络的中心节点,商河县是人口流动的边缘节点,如图 10 所示。同时,历城区还是整个网络图的桥节点,该市各区域夜间人口流动依靠历城区作为中转实现迁移。

4) 某市夜间活跃人口省内来源。从夜间活跃人口中的外省来源绝对数量可以看出,某市周边德州、泰安、济宁和菏泽来源最多(见图 11)。考虑外市常住人口基数后(即活跃人口/地区常住人口),发现某市周边泰安、德州两市和莱城区人口在夜间留驻该市的意愿最强。符合距离越近的市被吸引参与该市“夜经济”活动人口比例越高的特点。进一步分析,相近区域受该市夜经济

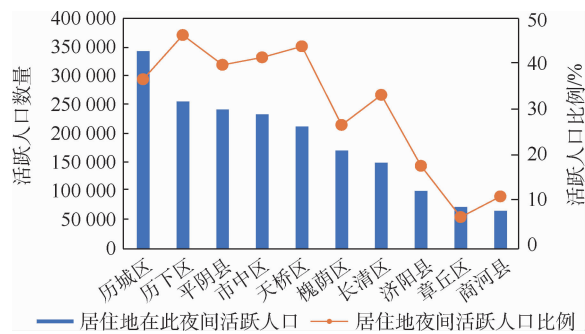


图 9 各区夜间活跃人口数量和比例

Fig. 9 Number and proportion of active population in different districts at night

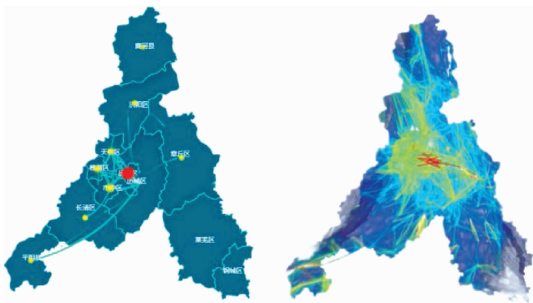
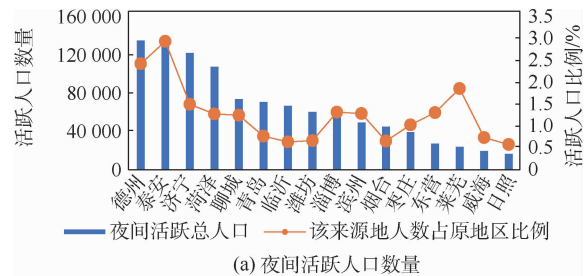
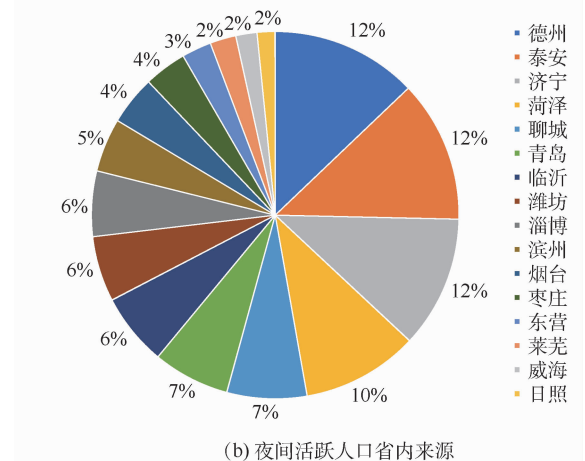


图 10 夜间人口迁移数量

Fig. 10 Population migration number at night



(a) 夜间活跃人口数量



(b) 夜间活跃人口省内来源

图 11 省内夜间活跃人口来源

Fig. 11 Sources of night active population in the province

辐射影响更强,距离较远区域受夜经济辐射能力相对较弱。地区人口基数来源:《山东省 2018 年统计年鉴》。

3.2.2 游在夜间

通过图 12 中展示的 4 个时间点人口分布来看,21:00 某市景点大明湖南岸和环护城河带的市民数量居多,说明夜经济政策点燃了市民夜间休闲的热情。

3.2.3 行在夜间

截止 2019 年 12 月 23 日,某市 24 小时公交 K101 已运行满 1 个月,夜间总运送人数超 2 000 人。图 13 为当日公交沿线热力图,路线串联了某市几大经济活跃区域(如芙蓉街、世贸购物中心、恒隆购物中心)和人口密集区域(如山大中心校区、裕园小区等),24 小时公交提供了便利的夜间回家途经,为该市夜经济注入了新活力。

3.2.4 玩在夜间

由图 8 可以看出,夜间活跃人口中 15~40 岁年龄段占比较高,年轻人常常利用互联网获取交通或消费信息,大数据时代让“线上”服务与“线下”O2O (online to offline) 模式被广泛应用。本节选取周六 19:00—24:00 区间,对某市人口使用地图导航和消费 APP 情况进行分析,如图 14 所示(列举部分 APP 展示)。当夜晚来临,居民活动总体趋势下降,使用支付宝、美团等消费 APP 的用户数量都高达上万。22:00 仍有万级人口使用各类交通导航工具,体现出某市的“夜经济”场景也是万物互联时代的夜经济场景。

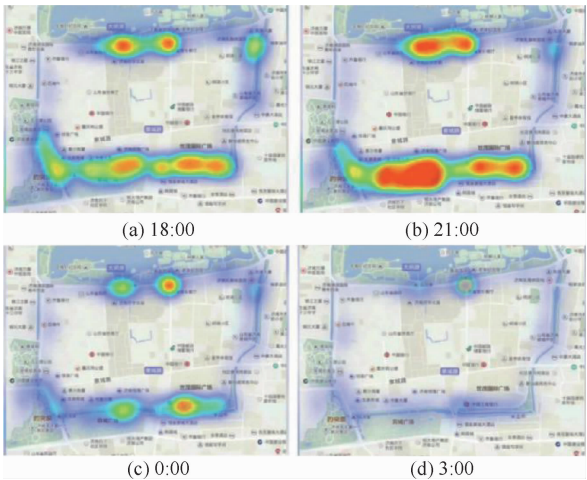


图 12 不同时间点的晚间热力图

Fig. 12 Thermal map of different time at night



图 13 K101 公交沿线热力图

Fig. 13 Thermal map along K101 bus

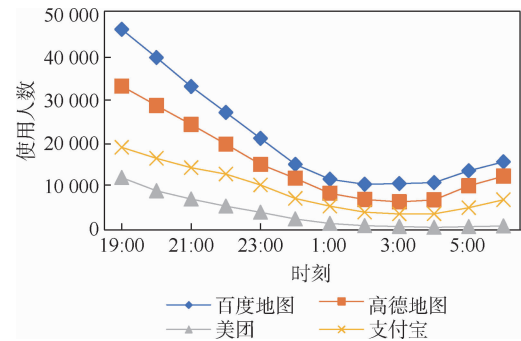


图 14 娱乐 APP 夜间使用人数

Fig. 14 Number of recreational APP users at night

3.2.5 政策评价和结论

采用基于 DID 的统计学方法,对夜间出游和夜经济政策进行相关性分析。由于政策影响夜间因素,选择是否执行政策和是否以夜间作为虚拟变量,以是否执行政策和是否为夜间的交互项作为政策评价系数,进行统计分析。

假设 1 实行夜经济政策对出游行为有显著正向影响,建立回归模型:

$$Y = \beta_0 + \beta_1 X_1 + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 (D_2 D_3)$$

(2)

式中: Y 为出游人数,为被解释变量; X_1 为人力最适宜温度(25℃)差值; D 为解释变量, D_1 表示是否为假期, D_2 表示是否为夜间, D_3 表示是否执行夜经济政策;根据 DID 理论, δ_3 为政策影响, δ_4 为政策实际效果,并考虑了夜间 D_2 与政策 D_3 的交互项。其中,采用分层回归算法:第 1 步,模型一只纳入 2 个解释变量对被解释变量进行解释;第 2 步,模型二利用其他解释变量对被解释变量进行解释。其分析结果如表 3 所示。

由表 3 可得,以 25℃ 为基准,当气温每偏离 1℃ 时,每小时出行人数降低 1 633 人;周末相较于非周末,每小时出行人数增加 12 974 人;晚上相较于白天,每小时外出休闲人数减少 9 066 人。将气温、节假日、昼夜等因素差分排除影响后,可得夜经济政策使得夜间每小时出行人数增加了 621 人。

表 3 分析结果

Table 3 Analysis results

变量符号	变量名称	模型一	模型二
β_0	截距	36 460.2	27 730.8 ***
β_1	温度		-1 633.7 *
δ_1	假期		12 974.2 **
δ_2	夜间	-9 066 *	-9 066.3 *
δ_3	政策影响	621.5 *	621'

注:“'”表示 $p < 0.1$,“*”表示 $p < 0.05$,“**”表示 $p < 0.01$,“***”表示 $p < 0.001$ 。

采用 DID 和层次回归方法对数据进行回归分析,结果表明,政策影响为正向且显著,通过了假设检验,证实了夜经济政策对夜间出游行为有着促进作用。夜间人们的活跃程度较白天显著性降低,该市夜经济还有更大的发展前景。

3.3 热点选址

选取 7 天内(2019 年 9 月 1 日至 7 日)不同时间段经济活跃区域景点的人流数、人口出游数、交通拥堵区域位置和基站附着点人流密度数据整合分析,并进行专家判断和趋势分析,确定位置在某市泉城路、大明湖、趵突泉、甸柳庄环联区域、万达广场等地点为活跃区域点。部分位置预测如图 15 所示,经济活跃区的判定有助于政府对夜经济发展做出决策,也有助于市民对夜市经济的响应,促进经济消费。

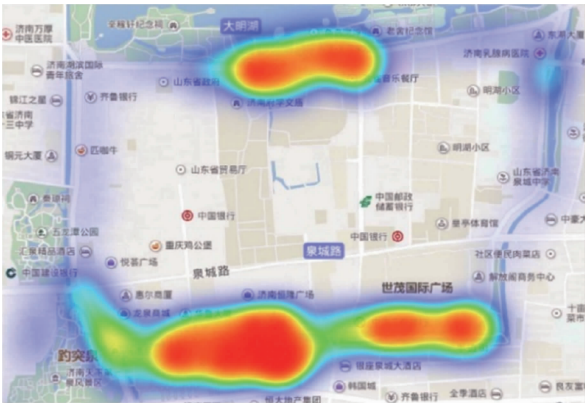


图 15 经济活跃区热力图

Fig. 15 Thermal map of economically active areas

4 结束语

本文结合反映人群活动特征的电信大数据开展研究,将其解释为轨迹数据和区域化数据进行分析应用:

- 1) 对电信大数据进行数据清洗和轨迹纠偏,定量抽取和分析关联特征,减少无效数据,提高数据质量。
- 2) 结合多源数据,对数据规则筛选及特征建模,多角度全方位分析用户行为,将时空区域数据进行可视化分析研究,采用双重差分统计模型评价区域经济政策。
- 3) 提出政策选址结论,用于指导城市热点商圈布局与调整,为区域经济发展提供决策依据。
- 本文还有很多方面可以改进,未来的研究重点集中在以下 3 个方面:

- 1) 在本文基础上结合应用信息推送、社交平台及微博热词推荐提取数据,多角度分析,进一步

增强信息可靠度。

2) 海量数据处理方面,采用多种数据处理算法融合进行清洗排序,增加数据质量可信度和可用度。

3) 将区域经济进一步扩展,从时间和空间延伸区域经济的可分析范围,例如,时间可选至重要节点日期等;空间可选购物、美食、旅馆、景区等。扩大经济区域效益范围,可为区域发展提供多方位参考。

参考文献 (References)

[1] SUN J B, YUAN J, WANG Y, et al. Exploring space-time structure of human mobility in urban space[J]. *Physica A: Statistical Mechanics and its Applications*, 2011, 390(5): 929-942.

[2] 罗名海, 秦思娟, 谭波, 等. 数据视角下的武汉市综合交通特征分析[J]. *地理空间信息*, 2020, 18(5): 1-7.

LUO M H, QIN S X, TAN B, et al. Characteristic analysis of the urban integrated transportation in Wuhan from the perspective of big data[J]. *Geospatial Information*, 2020, 18(5): 1-7 (in Chinese).

[3] 唐维红, 唐胜宏, 刘志华. 中国移动互联网发展报告(2020)[M]. 北京: 社会科学文献出版社, 2020.

TANG W H, TANG S H, LIU Z H. Annual report on China's mobile Internet development (2020)[M]. Beijing: Social Sciences Academic Press (China), 2020 (in Chinese).

[4] WANG H Y, CALABRESE F, DI LORENZO G, et al. Transportation mode inference from anonymized and aggregated mobile phone call detail records[C]//13th International IEEE Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2010: 318-323.

[5] AGUILERA V, ALLIO S, BENEZECH V, et al. Using cell phone data to measure quality of service and passenger flows of Paris transit system[J]. *Transportation Research Part C: Emerging Technologies*, 2014, 43(1): 198-211.

[6] 徐鹏飞. 基于夜间灯光遥感影像的城镇化时空特征研究[D]. 杭州: 浙江大学, 2019.

XU P F. The detection of spatio-temporal characteristics of urbanization based on the nighttime light data[D]. Hangzhou: Zhejiang University, 2019 (in Chinese).

[7] 张国亮, 朱瑞飞, 杜一博, 等. 吉林一号高分辨率夜光遥感影像在城市监测中的应用[J]. *卫星应用*, 2020(3): 27-33.

ZHANG G L, ZHU R F, DU Y B, et al. Application of Jilin-1 high-resolution luminous remote sensing images in urban monitoring[J]. *Satellite Application*, 2020(3): 27-33 (in Chinese).

[8] 海晓东, 刘云舒, 赵鹏军, 等. 基于手机信令数据的特大城市人口时空分布及其社会经济属性估测: 以北京市为例[J]. *北京大学学报(自然科学版)*, 2020, 56(3): 518-530.

HAI X D, LIU Y S, ZHAO P J, et al. Using mobile phone data to estimate the temporal-spatial distribution and socioeconomic attributes of population in megacities: A case study of Beijing[J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2020, 56(3): 518-530 (in Chinese).

[9] 刘华斌. 手机信令数据背景下城市交通出行方式选择辨识方法研究[D]. 北京: 北京交通大学, 2019.

LIU H B. Urban transportation modes recognition based on mobile signaling data[D]. Beijing: Beijing Jiaotong University, 2019 (in Chinese).

[10] 冀秉魁, 李杰, 姚雪萍, 等. 基于高德热力图的长春市停车规划分析[J]. *长春工程学院学报(自然科学版)*, 2018, 19(2): 29-31.

JI B K, LI J, YAO X P, et al. The urban parking planning analysis in Changchun city based on the Gaode thermodynamic diagram[J]. *Journal of Changchun Institute of Technology (Nature Science Edition)*, 2018, 19(2): 29-31 (in Chinese).

[11] 周艺华, 李广辉, 杨宇光, 等. 基于 GeoHash 的近邻查询位置隐私保护方法[J]. *计算机科学*, 2019, 46(8): 212-216.

ZHOU Y H, LI G H, YANG Y G, et al. Location privacy preserving nearest neighbor querying based on GeoHash[J]. *Computer Science*, 2019, 46(8): 212-216 (in Chinese).

[12] 史新颖, 夏元平, 毛曦, 等. DBSCAN 与 K-Means 相结合的手机大数据聚类方法研究[J]. *北京测绘*, 2019, 33(2): 132-137.

SHI X Y, XIA Y P, MAO X, et al. Research on mobile big data clustering method based on DBSCAN and K-Means[J]. *Beijing Surveying and Mapping*, 2019, 33(2): 132-137 (in Chinese).

[13] 杜翠凤, 余艺, 蒋超. 基于空间密度聚类的移动用户热点区域识别方法[J]. *移动通信*, 2015, 39(16): 40-43.

DU C F, YU Y, JIANG C. Mobile user hotspot recognition method based on space density clustering[J]. *Mobile Communications*, 2015, 39(16): 40-43 (in Chinese).

[14] 孙强. 空间 K-Means 挖掘算法在基于手机信令的 OD 分析中的应用[J]. *科技资讯*, 2018, 16(36): 14-16.

SUN Q. Application of spatial K-Means mining algorithm in OD analysis based on mobile phone signaling[J]. *Science and Technology Information*, 2018, 16(36): 14-16 (in Chinese).

[15] 苗壮. 基于手机信令数据的数据清洗挖掘与居民职住空间分析[D]. 成都: 西南交通大学, 2017.

MIAO Z. Research on data cleaning, mining, jobs and residential locations based on mobile phone signaling data[D]. Chengdu: Southwest Jiaotong University, 2017 (in Chinese).

[16] 余锦斌. 基于手机信令的数据分析引擎设计与实现[D]. 南京: 东南大学, 2018.

YU J B. Design and implementation of analysis engine based on mobile phone data[D]. Nanjing: Southeast University, 2018 (in Chinese).

[17] BIANCHINI F. Night cultures, night economies[J]. *Planning Practice and Research*, 1995, 10(2): 121-126.

[18] 吴敏. 河北省城市夜经济研究[D]. 天津: 河北工业大学, 2015.

WU M. Research on Hebei province urban night economy[D]. Tianjin: Hebei University of Technology, 2015 (in Chinese).

[19] ZHU X H, QIAN T N, WEI Y G. Do high-speed railways accelerate urban land expansion in China? A study based on the multi-stage difference-in-differences model[J]. *Socio-Economic Planning Sciences*, 2020, 71: 100846.

Application of space time regional economy visualization
based on telecom big data analysis

LI Na, LIU Wenmin^{*}, MENG Fanrui, LIU Yan

(Shandong Branch of National Computer Network Emergency Response Technical Team/Coordination Center of China, Jinan 250002, China)

Abstract: Currently, the number of mobile phone users in China has reached 1.59 billion. Under the huge population base, the telecom big data characteristics reflect the characteristics of crowd activities to a certain extent, which can reflect the development status of specific regions. The application of space time regional economy visualization processes and extracts the information from massive telecom big data by data mining technology to improve data quality and screens the data in different rules and extracts features by modeling techniques of the data. The data combined with multi-source information, such as electronic map data and traffic data are used to analyze user behavior characteristics from multiple perspectives. The application analysis makes use of the data to visualize and research the space time regional economic situation and analyze life attributes of the residents. At the same time, use the difference-in-differences (DID) model to evaluate regional economic policies. Based on the results of feature analysis, it can provide the decision-making basis for the location of regional economic development and guiding layout of urban business districts, improve the efficiency of urban system operation and expand the range of economic regional benefits.

Keywords: telecom big data; mobile phone signaling; data mining; regional economy; visualization

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0400

一种基于深度学习的恶意代码克隆检测技术

沈元¹, 严寒冰^{2,*}, 夏春和¹, 韩志辉²

(1. 北京航空航天大学 计算机学院, 北京 100083; 2. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘 要: 恶意代码克隆检测已经成为恶意代码同源分析及高级持续性威胁 (APT) 攻击溯源的有效方式。从公共威胁情报中收集了不同 APT 组织的样本, 并提出了一种基于深度学习的恶意代码克隆检测框架, 目的是检测新发现的恶意代码中的函数与已知 APT 组织资源库中的恶意代码的相似性, 以此高效地对恶意软件进行分析, 进而快速判别 APT 攻击来源。通过反汇编技术对恶意代码进行静态分析, 并利用其关键系统函数调用图及反汇编代码作为该恶意代码的特征。根据神经网络模型对 APT 组织资源库中的恶意代码进行分类。通过广泛评估和与 MCrab 模型的对比可知, 改进模型优于 MCrab 模型, 可以有效地进行恶意代码克隆检测与分类, 且获得了较高的检测率。

关 键 词: 深度学习; 高级持续性威胁 (APT) 组织; 克隆检测; 控制流图 (CFG); 系统函数调用图

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965 (2022) 02-0282-09

克隆检测主要包括源代码检测及二进制代码检测, 广泛应用于漏洞发现、代码克隆检测、用户端崩溃分析等。随着科技的日益发展, 大量的物联网设备投入使用, 恶意代码分析变得越来越重要。据 Gartner 分析在 2017 年时全球已经有 84 亿物联网设备投入使用, 比 2016 年增长 31%, 2020 年达到 204 亿^[1]。而物联网的快速发展, 导致各种网络攻击及恶意代码也随之增多。因此, 恶意代码分析变得十分迫切。

同时, 由于针对企业或政治动机的高级持续性威胁 (APT) 攻击的快速增长, APT 攻击已经成为网络安全领域的热门话题, 这需要安全从业者快速分析恶意软件之间的同源性来确定 APT 攻击背后的群体。当获得新的恶意代码时, 试图检测其是否是通过其他恶意代码的源代码进行编译得到的^[2]。众所周知, 同家族样本或同类型样本

之间存在大量共享代码库。

但是出于不同目的, APT 组织包含不同族和不同类型的恶意代码, 如 Duqu、Stuxnet 和 Flame 属于同一个 APT 组织^[3-7]。大多数恶意软件研究人员致力于分析、检测、家族分类等, 但这些努力无法从根本上解决恶意软件的同源性识别问题。因此, 安全从业者需要更加关注恶意软件的同源性分析, 并发现不同恶意软件之间的同源性, 进而挖掘攻击背后的 APT 组织。

学术界通过利用恶意代码克隆检测来解决恶意软件同源性问题。因此, 为了满足处理大样本的需求, 需要提出一种能够自动化进行的恶意软件克隆检测技术, 从而识别恶意软件的同源性, 大量减少了人工分析的工作量, 提高了效率。

本文通过对每个恶意软件样本进行静态反汇编分析, 再根据函数的控制流图 (CFG) 构建

收稿日期: 2020-08-09; 录用日期: 2021-01-21; 网络出版时间: 2021-03-17 17:20
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210316.1557.005.html
基金项目: 国家自然科学基金 (U1736218); 国家科技重大专项 (2018YFB0804704); 北航青年拔尖人才支持计划 (YWF-20-BJ-J-1038)
* 通信作者。E-mail: yhb@cert.org.cn
引用格式: 沈元, 严寒冰, 夏春和, 等. 一种基于深度学习的恶意代码克隆检测技术[J]. 北京航空航天大学学报, 2022, 48 (2): 282-290. SHEN Y, YAN H B, XIA C H, et al. Malicious code clone detection technology based on deep learning [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48 (2): 282-290 (in Chinese).

其自定义函数的反汇编代码文本,以及整个样本的系统函数调用图(AFCG)做为恶意软件的特征进行提取并向量化,换句话说,将恶意软件建模成一种语言,其可行性已经由 Awad 等^[8]验证。构建了一个神经网络模型来对数据样本集进行分类。实验表明,本文方法可以有效地进行恶意代码克隆检测与分类,且获得了较高的准确率和召回率。

同时,本文方法已经证明了自动实现恶意软件克隆检测有助于帮助恶意软件分析师和决策者做出更快的判断,并可以作为自动分析的第一步。

本文的主要贡献如下:

1) 将公共威胁情报中的样本集分类到不同的 APT 组织,并构建 APT 组织的样本数据库,其中提取了约 6 900 个样本和约 80 万个自定义函数。

2) 提出了一种新方法,首先对每个恶意样本进行静态反汇编分析,根据样本函数的 CFG 构建其自定义函数的反汇编代码文本,以及整个样本的 AFCG 为恶意软件的特征。然后,利用之前构建的神经网络模型 CNN-SLSTM,将卷积神经网络(CNN)模型与改进的递归神经网络长期短期记忆(LSTM)模型相结合,对 APT 组织样本进行分类。

3) 通过广泛评估,同时与 MCrab 模型^[9]进行了对比表明,改进模型(GMCrab)优于 MCrab 模型,可以有效地进行恶意代码克隆检测与分类,且获得了较高的检测率。

1 相关工作

代码克隆检测技术有很多种方法,如传统方法中的静态抄袭检测或克隆检测,包括基于字符串的检测方法^[10-12]、基于令牌的检测方法^[13-15]、基于语法树的检测方法^[16-18]和基于程序依赖图(PDG)的检测方法^[19-22]。有些方法是基于样本的源代码进行检测,不适用于实践中,尤其不适用于恶意代码。在基于动态标记的方法中,Tamada 等^[1]提出为 Windows 应用程序进行 API 标记;Schuler 等^[23]提出为 Java 程序进行动态标记。但是,由于它们依赖于动态分析,很难将其扩展到其他架构并适应嵌入式设备,而且在动态分析中的代码覆盖率问题是一个固有的挑战。

由于研究证明了将机器学习和深度学习技术应用于代码分析的有用性,White 等^[24]进一步提出了用 DeepRepair 来检测源代码片段之间的相

似性。Mou 等^[25]提出了基于程序抽象语法树和基于树的 CNN 来对源代码片段进行克隆检测。Liu 等^[26]使用了其代码特征调用来表征二进制函数。针对不同架构的相似性检测,Feng 等^[27-28]提出了一种跨架构二进制相似性检测的解决方法,分别利用传统机器学习和深度学习将函数的 CFG 转换为用于相似性比较的向量。文献[29]引入了一种选择性内联技术来捕获函数语义,并提取了不同长度的部分踪迹来对函数建模。

虽然恶意代码属于二进制代码中的一种,但因其自身的独特性,并不能完全使用于上述的检测方法。因此在本文中,通过参考上述方法,提出了一种通过利用恶意代码进行反汇提取的新特征来构建恶意代码的克隆检测特征向量,其主要利用恶意软件 AFCG 及自定义函数的语义特征并结合深度学习技术,通过构建神经网络模型对恶意代码进行克隆检测并对 APT 组织进行分类。

2 本文方法

2.1 系统模型

本文提出了一种基于深度学习的恶意代码克隆检测技术,该技术可以快速、有效地检测出恶意代码之间的相似性,识别出该恶意代码相关的 APT 组织类别信息。

本文原型系统模型如图 1 所示。包括以下主要步骤:①APT 样品采集;②原始特征提取;③利用 Siamese 网络对特征进行清洗,由于一些恶意代码应用了较多的开源库,下载了大量开源库,并使用局部敏感哈希算法通过匹配去除大量公共函数;④分类模型。其中,第①步旨在从公共威胁情报中收集 APT 攻击的样本。第②步利用反汇编技术提取样本的 AFCG 及自定义反汇编代码作为样本的原始特征。第③步进行特征向量化并用作为模型的输入。第④步构建一个神经网络模型来对带有标记的 APT 组织样本进行分类。系统模型的使用场景如图 2 所示。通过对已有的知识库进行训练,并建立模型,当发生新的 APT 攻击时,病毒引擎将首次分析样本。当没有强大的 IoC(indicators of compromise)信息(如 IP 地址、域名、URL 等)来确定 APT 攻击背后的组织时,可以使用该系统来预测和判断攻击背后与之相关的 APT 组织,并进一步指导领域专家做出判别。

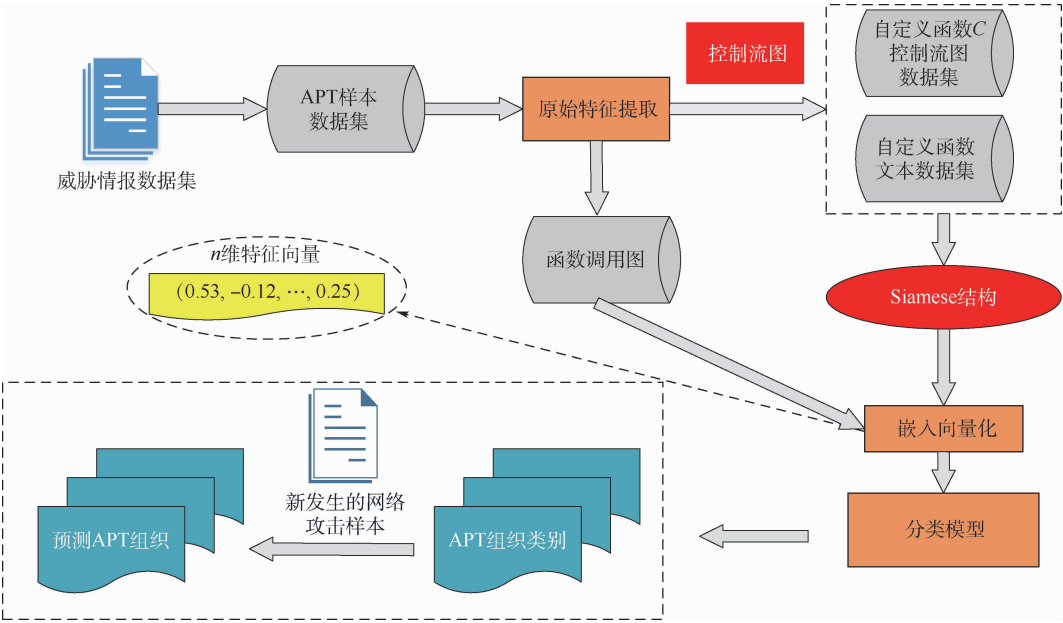


图 1 原型系统 GMCrab 工作流程

Fig. 1 Prototype system GMCrab workflow

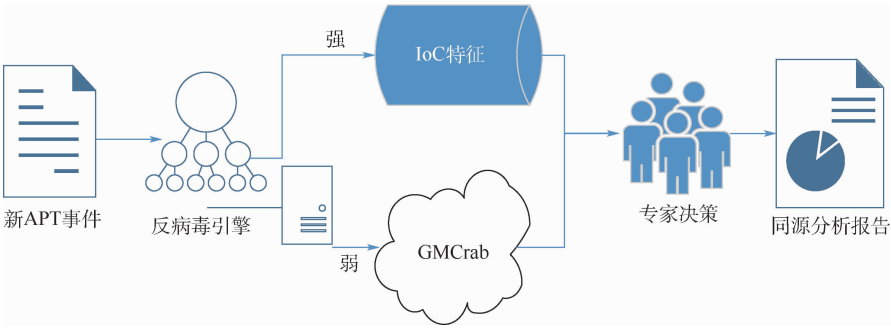


图 2 GMCrab 部署图

Fig. 2 Deployment of GMCrab

2.2 恶意代码特征提取

无论是对恶意代码进行检测还是分类,恶意代码的特征提取在整个工作流程中都起着至关重要的作用。

1) 恶意代码文本特征。CFG 是二进制分析中常用的特征。因此,根据自定义函数的 CFG,先通过利用 IDA Pro 进行反汇编构成 CFG,将非循环路径上每 3 个代码块中的反汇编代码构成 1 句长语句。根据 CFG 的执行路径,可以将图 3 分解为 3 个块来获得一系列块,如式(1)所示。使用 3 个序列块中的反汇编代码形成 1 个块表示该自定义函数的文本。最终对构造的文本进行规范化,并统一标记变量名称和内存地址,以提高系统模型在后续分类任务中的准确率和召回率。

(1,2,3),(1,2,4),(2,4,5)

2) 恶意代码 AFCG。在本文中,系统函数指的是调用操作系统提供的 API 函数,如 Window 系统的 MSDN 中的 API。图 4 为一个关键函数的

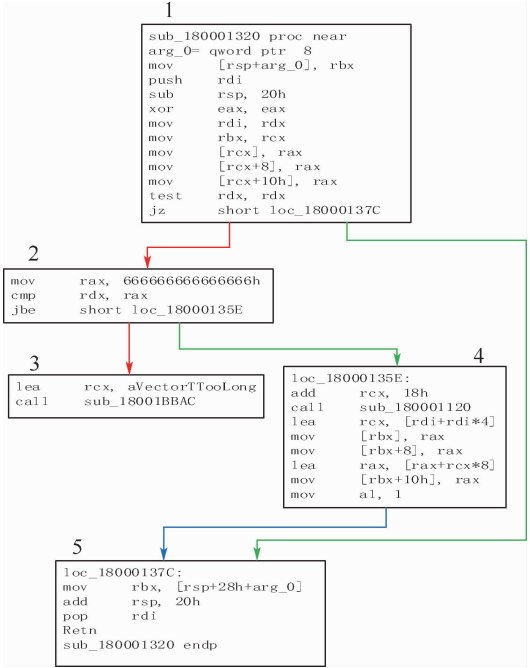


图 3 CFG 样例

Fig. 3 CFG example

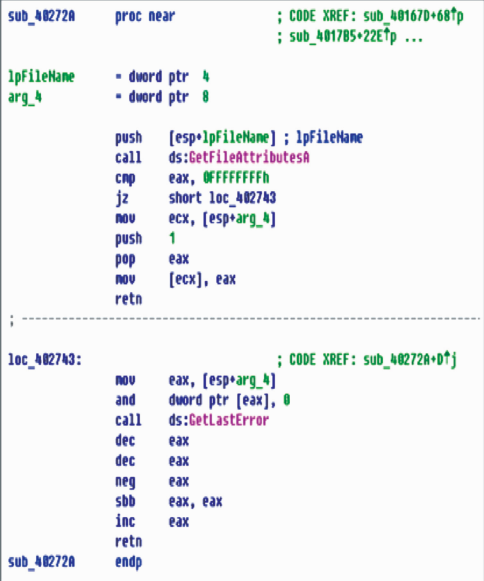


图 4 系统函数调用示例

Fig.4 Example of system function call

示例,该函数调用了 2 个 API,其 API 序列为 Get-FileAttributesA→GetLastError。函数调用图表示样本中函数间的调用关系。由于需要使用AFCG,但是非系统调用函数在整个样本的占比很高,导致系统函数间可能不存在直接的调用关系,需要通过给图的边加入权值来代替图中的非系统调用函数,即AFCG(在函数调用图的基础上将非系统调用的函数去除压缩而成)。将有向无权的函数

调用图压缩为有向带权值的 AFCG,其权值的赋值遵循以下 4 个原则:

- ① 若 2 个系统函数间存在直接调用关系,则权值为 1。
- ② 若 2 个系统函数间不存在直接调用关系,则权值为经过的非系统函数个数加 1。
- ③ 若 2 个系统函数间存在多条调用路径时,权值为最短路径跳数。
- ④ 若 2 个系统函数间存在的调用路径都经过另外 1 个系统函数,则这 2 个系统函数间不存在直接的有向带权值路径。

3) 数据清洗。收集了 3 000 个良性样本,同样利用反汇编技术收集了约 2 万个自定义函数的代码片段。利用对等网络架构,通过计算代码片段的局部敏感哈希值进行相似性计算,从而去除与其相似的恶意代码库中提取的函数。最终利用剩下的函数库作为输入特征进行分类。

3 基于深度学习克隆检测算法

本文模型沿用文献[9]中的 CNN-SLSTM 模型,该模型的结构如图 5 所示,其由 2 个部分组成:①C-LSTM 模型中的 CNN 模型;②改进的 SLSTM 模型,即引入了结构特征。详细的模型介绍如下:

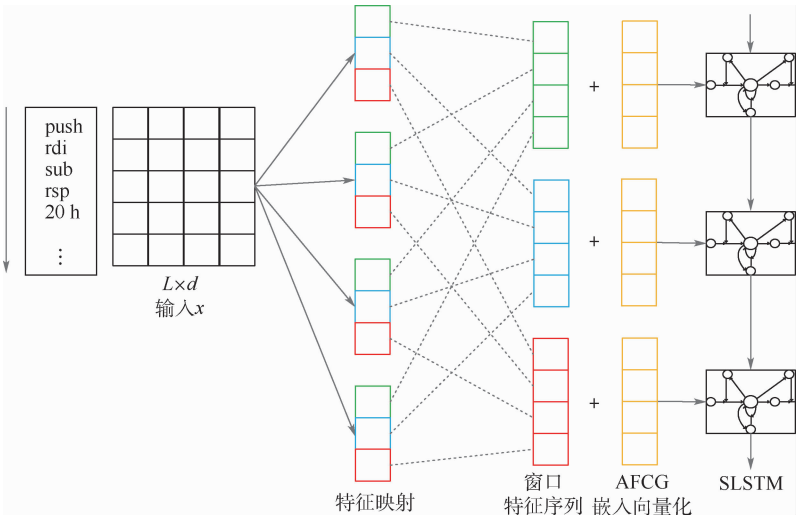


图 5 语义模型的 CNN-SLSTM 的体系结构

Fig.5 Architecture of CNN-SLSTM for semantic model

1) 通过 CNN 进行特征提取。对构成的文本特征进行建模,CNN 利用卷积滤波器提取句子不同位置上的 n-gram 特征,该方法表现出色。文献[30]描述的 n-gram 特征提取方法即对从样本中提取的 CFG 反汇编代码文本执行相同的处理。

对于文本中的每个位置 i ,都有 1 个带有 k 个连续单词向量的窗口向量 $\mathbf{X}_i$,表示为

$$\mathbf{X}_i = [x_i, x_{i+1}, \dots, x_{i+k-1}] \tag{2}$$

式中: $x_i \in \mathbf{R}^d$ 为反汇编代码文本中第 i 个字的 d 维子向量; k 为滤波器的长度。

滤波器 w 用于生成新的特征映射,公式如下:

$$c \in \mathbf{R}^{L-k+1} \tag{3}$$

式中: L 为文本的长度。

特征映射的每个元素 c_i 的产生如下:

$$c_i = f(w \cdot X_i + b) \tag{4}$$

式中: b 为一个偏置项, $b \in \mathbf{R}$; f 为非线性函数(如 sigmoid、双曲正切等); w 为权值。

选择 ReLU 作为本文实验的滤波器,通过把滤波器应用于每个窗口 x_i 以产生特征图。

$$\mathbf{C} = [c_1, c_2, \dots, c_n] \tag{5}$$

式中: c_i 为使用第 i 个滤波器生成的特征映射;每行 $\mathbf{C} \in \mathbf{R}^{(L-k+1) \times n}$, C_i 为从位置 i 处的窗口矢量的 n 个滤波器生成的新特征表示。

在这些处理步骤之后,可以得到 1 个高级窗口表示,其将被用作下 1 个 SLSTM 模型的输入之一。本文只进行卷积而不在特征映射上应用 max-over-time 池化操作来获取最大值。这是因为在 LSTM 被指定用于序列输入,如果使用池化,则会破坏由 CNN 提取的高级窗口特征。

2) SLSTM 网络。众所周知, RNN^[31-32] 中的问题在于:在训练期间,梯度向量的分量可以在长序列上呈指数增长或衰减^[33]。梯度爆炸或消失的问题使得 RNN 模型难以学习序列中的长距离相关性,而已经提出的 LSTM 模型可以解决特定学习长期依赖性的问题。本文结合恶意代码源性研究问题,将对应于样本的系统函数调用图结构特征引入标准 LSTM 模型,并称之为 SLSTM。

如图 5 所示, SLSTM 模型的输入由 2 个部分组成:①CNN 提取的高级窗口表示;②AFCG 的嵌入表示。同一个样本中的每个自定义函数的 CFG 对应相同 AFCG,即每个自定义函数的反汇编代码文本对应同样的 AFCG 结构特征。在图 5 特征映射层和窗口特征序列层中,相同颜色的块对应于同一窗口特征,虚线将窗口的特征与源特征图连接起来,并且黄色块表示与样本对应的 AFCG 嵌入向量,其与 CNN 模型学习的窗口特征一起被输入到 SLSTM 模型中。整个模型的最终输出的是 SLSTM 的隐藏单元。

参考 Dai 等^[34] 提出的图形嵌入网络,将 AFCG 表示为 $g = \langle V, E \rangle$, 其中 V 和 E 分别为顶点和边的集合。此外,图中的每条边 e 都具有附加特征 x_e , 对应 AFCG 中的边的权值。图形嵌入网络将先计算每个顶点 $v \in V$ 的 p 维特征 μ_v , 再将 g 的嵌入向量 μ_g 计算定义为这些顶点的聚合,公

式为

$$\mu_g := \sum v \in V^{(\mu_v)} \tag{6}$$

将 μ_g 表示为结构向量 S , 将由 c 维图形嵌入和由 CNN 模型获得的高级窗口表示生成的新连续表示馈送到 SLSTM 模型中。

在每个时间将 SLSTM 单元定义为 $\mathbf{R}^d$ 中的向量集合:输入门 i_t 、遗忘门 f_t 、输出门 o_t 、存储器单元 c_t 和隐藏状态 h_t 。在模型中,将结构向量 S 添加到输入门、遗忘门、单元和输出门。改进后的 LSTM 模型公式如下:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i + P_i S) \tag{7}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f + P_i S) \tag{8}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_{t-1} + b_o + P_i S) \tag{9}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c + P_i S) \tag{10}$$

$$c_t = f_t \odot C_{t-1} + i_t \odot \tilde{c}_t \tag{11}$$

$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

式中: b 为偏差项; d 为 SLSTM 单位的数量; x_t 为当前时间步的输入; σ 为逻辑 sigmoid 函数,其输出为 $[0, 1]$; $\odot$ 表示数组元素依次相乘; $P_i S$ 是对原始 LSTM 等式的修改。

遗忘门控制擦除存储器单元的每个单元的量,输入门控制每个单元的更新量,输出门控制内部存储器状态的曝光。SLSTM 可以处理时间序列数据以学习长期依赖关系,因此,选择其作为 CNN 模型的下一个输入模型来学习代码段中的依赖关系。最终在 SLSTM 模型的最后 1 个步骤中将 softmax 层添加到隐藏状态的输出以进行分类。

4 实验与结果

4.1 数据集

本文利用 2 年时间收集了约 8 000 例不同 APT 组织样本,其中大部分来源于公共威胁情报,利用情报中的样本哈希值,通过 VirusTotal 进行检索与筛选,获得与该组织对应的样本文件。基于大量的公共威胁情报,构建了 APT 组织的样本数据库,并从 10 个 APT 组织的 6 972 个样本中提取了 680 718 个自定义函数。将这些自定义函数形成数据集并进一步以 8:2 的比例随机分为训练数据集和测试数据集,还使用了 10% 的训练集进行交叉验证,相关统计数据如表 1 所示。

4.2 实验环境

在 Python 中使用 TensorFlow 实现了本文的模型。利用 GPU 进行模型训练。在文本表示中,使用 word2vec 向量初始化单词向量。在最终设

表 1 APT 组织样本数据集
Table 1 APT group samples dataset

APT 组织	样本个数	函数个数	训练集	验证集	测试集	标签
patchwork	964	83 230	62 430	8 323	20 800	0
hangover	768	75 636	56 727	7 564	18 909	1
stuxnet	829	79 624	59 718	7 962	19 906	2
darkhotel	658	53 290	39 968	5 329	13 322	3
lazarus	721	65 997	49497	6 560	16 500	4
apt28	538	62 548	46 911	6 255	15 637	5
deeppanda	682	75 629	56 722	7 563	18 907	6
apt10	721	85 476	64 107	8 548	21 369	7
gaza	530	48 926	36 695	4 893	12 231	8
turla	561	50 362	37 772	5 036	12 590	9
总计	6 972	680 718	510 547	68 033	170 171	

置中,使用了整数线性单位,过滤窗口设置为 3、4 和 5,每个过滤窗口得到 100 个特征图,SLSTM 的内存尺寸为 100,随机失活率为 0.5,mini-batch 设为 50,这些值是通过开发集上的网格搜索选择的。本文的模型在单词序列馈送到卷积层之前,将 dropout 应用于单词向量,并在 softmax 层之前将其输入到 LSTM 模型的输出。使用 Adadelata 更新规则,通过随机梯度下降在乱序的 mini-batch 上进行训练,同时还在 softmax 层的权重中添加了因子为 0.001 的 L2 正则化。

4.3 模型比较

选择 MCrab 模型作为比较的基线,2 个模型的对比结果如图 6 所示。此外,还选择了一些比较先进的文本分类方法进行比较,如表 2 所示。将本文模型与其他模型进行比较,发现 CNN-SLSTM 模型优于其他模型。总的来说,本文模型显示出强大的竞争力。这说明 CNN-SLSTM 模型可以学习 AFCG 的结构特征和其反汇编代码的文本特征。AFCG 的结构特征在恶意代码克隆检测问题中具有重要意义。此外,C-LSTM模型的性能优于单个 CNN 模型和单个 LSTM 模型,这意味着使用 CNN 模型提取高级文本表示在学习跨序列的长期依赖性中起着重要作用。在 CNN 模型中,将其与来自 word2vec 的预训练向量、微调词向量和多通道进行比较,具有动态 k-max 汇集的 DCNN 模型^[35]。对于 LSTM 模型,将其与 2 个树状结构的 LSTM 模型进行了比较^[36],同时也与 LSTM 模型和 Bi-LSTM 模型进行了比较。此外,将改进的 GMCrab 模型与 MCrab 模型进行了对比,发现在准确率上略优于 MCrab 模型,原因在于:在恶意样本特征提取后做了较好的特征清洗及新特征的选择,提高了模型准确率。

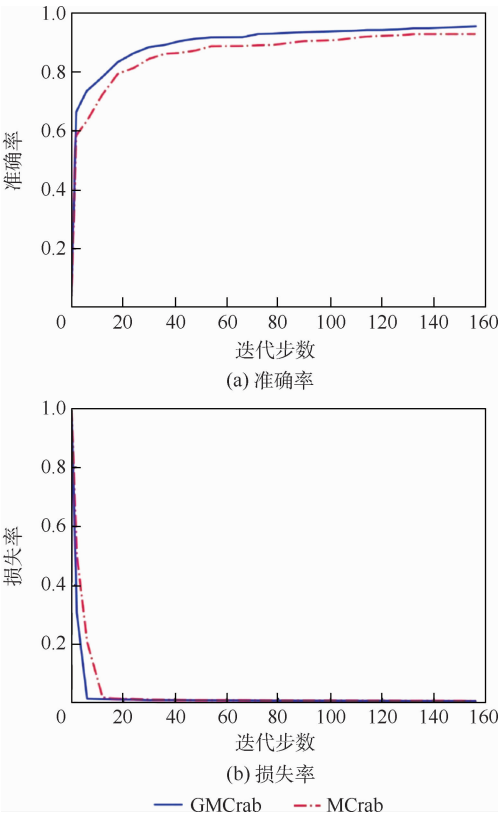


图 6 不同模型的准确率和损失率对比
Fig. 6 Comparison of accuracy and loss for different models

4.4 模型分析

本节从以下 3 个方面评估本文模型:①卷积层中不同滤波器配置对模型性能的影响;②数据量对模型性能的影响;③模型在恶意软件同源性分析中的可行性。

1) 过滤器大小。在本文模型中,使用 CNN 模型通过过滤器捕获本地 n-gram 文本特征。在 2 种情况下进行了滤波器长度为 3、4 和 5 的研究:具有相同滤波器长度的单个卷积层和多个卷积层,并行不同长度的滤波器。对于第 1 种情况,每个 n-gram 窗口在卷积之后被转换成卷积特征,

表 2 不同模型比较

Table 2 Comparison with different models

模型	准确率
CNN-static	0.836
CNN-non-static	0.857
CNN-multichannel	0.879
DCNN	0.886
LSTM	0.902
Bi-LSTM	0.881
C-LSTM	0.910
CNN-SLSTM	0.953

并且窗口表示的序列被馈送到 SLSTM 模型中。第 2 种情况使用了不同尺寸的滤波器,因此,在每个卷积层之后获得的窗口数量是不同的。根据最大滤波器长度切割窗口序列来获得最小窗口数。最终每个窗口由来自不同卷积层输出的串联表示。

2) 数据大小。为了进一步检查本文方法在不同大小数据集上的性能,以 1:100 的采样比率为每个 APT 组织选择样本,并将数据集分成 2 个不同大小的数据集。基于上述模型评估了不同大小的数据集,发现 SVM 在小数据集上的表现优于本文方法。

本节通过介绍数据集、实验环境,以及对实验进行评估,表明了与其他分类器相比,本文方法更具有优势。

5 结 论

本文利用 2 年时间收集了多达 8 000 例不同 APT 组织样本,同时提出了一种新方法,通过对每个恶意软件样本进行静态反汇编分析,再根据函数的 CFG 构建其自定义函数的反汇编代码文本,以及整个样本的 AFCG 作为恶意软件的特征相结合,利用神经网络模型(CNN-SLSTM)对 APT 组织样本进行分类。

通过广泛评估并与 MCrab 模型进行对比,结果表明,改进 GMCrab 模型优于 MCrab 模型,可以有效地进行恶意代码克隆检测与分类,且获得了较高的检测率。但是,一些 APT 小组恶意模仿其他 APT 组织的编码风格,会影响模型判断,以后将努力解决这个问题。此外,将扩展数据集添加更多 APT 组织样本,并更全面地进行恶意软件克隆检测分析。

参考文献 (References)

[1] TAMADA H,OKAMOTO K,NAKAMURA M,et al. Dynamic soft-

ware birthmarks to detect the theft of Windows applications[C]// International Symposium on Future Software Technology,2004.

[2] PFEFFER A,CALL C,CHAMBERLAIN J,et al. Malware analysis and attribution using genetic information[C]//7th International Conference on Malicious and Unwanted Software. Piscataway:IEEE Press,2012:39-45.

[3] RUTTENBERG B,MILES C,KELLOGG L,et al. Identifying shared software components to support malware forensics[C]// International Conference on Detection of Intrusions, Malware and Vulnerability Assessment. Berlin:Springer,2014:21-40.

[4] BENCÁSÁTH B,PÉK G,BUTTYÁN L,et al. The cousins of stuxnet:Duqu,Flame and Gauss[J]. Future Internet,2012,4 (4):971-1003.

[5] BENCÁSÁTH B,PÉK G,BUTTYÁN L,et al. Duqu: A Stuxnet-like malware found in the wild[R]. Hungary: CrySyS Lab Technical Report,2011,14:1-60.

[6] GOSTEV A,KUZNETSOV I. Stuxnet/Duqu: The evolution of drivers[EB/OL]. (2011-12-28) [2020-08-01]. [http://www.securelist.com/en/analysis/204792208/Stuxnet Duqu The Evolution of Drivers](http://www.securelist.com/en/analysis/204792208/Stuxnet_Duqu_The_Evolution_of_Drivers).

[7] CHIEN E,OMURCHU L,FALLIERE N. Duqu: The precursor to the next Stuxnet[C]// LEET 12: Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats,2012.

[8] AWAD Y,NASSAR M,SAFA H. Modeling malware as a language[C]//IEEE International Conference on Communications (ICC). Piscataway:IEEE Press,2018:1-6.

[9] LIU J R,SHEN Y,YAN H B. Functions-based CFG embedding for malware homology analysis[C]//26th International Conference on Telecommunications (ICT). Piscataway:IEEE Press,2019:220-226.

[10] BAKER B S. On finding duplication and near-duplication in large software systems[C]//Proceedings of 2nd Working Conference on Reverse Engineering. Piscataway:IEEE Press,1995:86-95.

[11] BILENKO M,MOONEY R J. Adaptive duplicate detection using learnable string similarity measures[C]//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD' 03. New York:ACM,2003:39-48.

[12] BROMLEY J,BENTZ J W,BOTTOU L,et al. Signature verification using a "Siamese" time delay neural network[J]. International Journal of Pattern Recognition and Artificial Intelligence,1993,7(4):669-688.

[13] DEERWESTER S,DUMAIS S T,FURNAS G W,et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science,1990,41(6):391-407.

[14] KAMIYA T,KUSUMOTO S,INOUE K. CCFinder: A multilingualistic token-based code clone detection system for large scale source code[J]. IEEE Transactions on Software Engineering,2002,28(7):654-670.

[15] SCHLEIMER S,WILKERSON D S,AIKEN A. Winnowing: Local algorithms for document fingerprinting[C]//Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data-SIGMOD' 03. New York:ACM,2003:76-85.

- [16] PRECHELT L, MALPOHL G, PHILIPPSEN M. Finding plagiarisms among a set of programs with JPlag[J]. Journal of Universal Computer Science, 2002, 8(11): 1016-1038.
- [17] JIANG L X, MISHERGHI G, SU Z D, et al. Deckard: Scalable and accurate tree-based detection of code clones[C]//29th International Conference on Software Engineering. Piscataway: IEEE Press, 2007: 96-105.
- [18] KOSCHKE R, FALKE R, FRENZEL P. Clone detection using abstract syntax suffix trees[C]//Proceedings of 13th Working Conference on Reverse Engineering. Piscataway: IEEE Press, 2006: 253-262.
- [19] PEWNY J, SCHUSTER F, BERNHARD L, et al. Leveraging semantic signatures for bug search in binary programs[C]//Proceedings of the 30th Annual Computer Security Applications Conference. New York: ACM, 2014.
- [20] GABEL M, JIANG L, SU Z. Scalable detection of semantic clones[C]//30th International Conference on Software Engineering. Piscataway: IEEE Press, 2008: 321-330.
- [21] LIU C, CHEN C, HAN J W, et al. GPLAG: Detection of software plagiarism by program dependence graph analysis[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD'06. New York: ACM, 2006: 872-881.
- [22] CRUSSELL J, GIBLER C, CHEN H. Attack of the clones: Detecting cloned applications on Android markets[C]//Computer Security-ESORICS. Berlin: Springer, 2012: 37-54.
- [23] SCHULER D, DALLMEIER V, LINDIG C. A dynamic birth-mark for Java[C]//22nd ACM International Conference on Automated Software Engineering. New York: ACM, 2007: 274-283.
- [24] WHITE M, TUFANO M, MARTINEZ M, et al. Sorting and transforming program repair ingredients via deep learning code similarities[C]//26th International Conference on Software Analysis, Evolution and Reengineering. Piscataway: IEEE Press, 2019: 479-490.
- [25] MOU L, LI G, ZHANG L, et al. Convolutional neural networks over tree structures for programming language processing[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 1287-1293.
- [26] LIU B C, HUO W, ZHANG C, et al. α Diff: Cross-version binary code similarity detection with DNN[C]//33th ACM/IEEE International Conference on Automated Software Engineering. New York: ACM, 2018: 667-678.
- [27] FENG Q, ZHOU R D, XU C C, et al. Scalable graph-based bug search for firmware images[C]//Proceedings of ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 480-491.
- [28] XU X J, LIU C, FENG Q, et al. Neural network-based graph embedding for cross-platform binary code similarity detection[C]//Proceedings of ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2017: 363-376.
- [29] CHANDRAMOHAN M, XUE Y X, XU Z Z, et al. BinGo: Cross-architecture cross-OS binary search[C]//24th ACM SIGSOFT International Symposium. New York: ACM, 2016: 678-689.
- [30] ZHOU C T, SUN C L, LIU Z Y, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.
- [31] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [32] GHOSH S, VINYALS O, STROPE B, et al. Contextual LSTM (CLSTM) models for large scale NLP tasks[EB/OL]. (2016-05-31)[2020-08-01]. <https://arxiv.org/abs/1602.06291>.
- [33] KOLEN J F, KREMER S C. Gradient flow in recurrent nets: The difficulty of learning longterm dependencies[J]. A Field Guide to Dynamical Recurrent Networks, 2001, 28(2): 237-243.
- [34] DAI H J, DAI B, SONG L. Discriminative embeddings of latent variable models for structured data[C]//Proceedings of the 33rd International Conference on Machine Learning. New York: ACM, 2016: 2702-2711.
- [35] KALCHBRENNER N, GREFFENSTETTE E, BLUNSOM P. A convolutional neural network for modelling sentences[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 655-665.
- [36] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015: 1556-1566.

Malicious code clone detection technology based on deep learning

SHEN Yuan¹, YAN Hanbing^{2,*}, XIA Chunhe¹, HAN Zhihui²

(1. School of Computer Science and Engineering, Beihang University, Beijing 100083, China;

2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: Malicious code clone detection has become an effective way to analyze malicious code homology and advanced persistent threat (APT) attacks. In this paper, we collect samples of different APT organizations from public threat intelligence, and propose a deep learning based malicious code clone detection framework to detect the similarity between the functions in newly discovered malicious code and the malicious code in known APT organizational resources in order to efficiently analyze malware and quickly identify the source of APT attacks. We perform static analysis of malicious code through disassembly technology, use its key function call graph and disassembly code as the features of the malicious code, and then classify the malicious code in the APT organization library according to the neural network model. Through extensive evaluation and comparison with our previous models (MCrab), the improved model is better than the previous model, which can effectively detect and classify malicious code clones and obtain higher detection rate.

Keywords: deep learning; advanced persistent threat (APT) groups; clone detection; control flow graph (CFG); system function call graph

Received: 2020-08-09; **Accepted:** 2021-01-21; **Published online:** 2021-03-17 17:20
URL: kns.cnki.net/kcms/detail/11.2625.V.20210316.1557.005.html
Foundation items: National Natural Science Foundation of China (U1736218); National Science and Technology Major Project (2018YFB0804704); Beihang Youth Top Talent Support Program (YWF-20-BJ-J-1038)
*** Corresponding author.** E-mail: yhb@cert.org.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0390

TLS 密码套件的流量数据随机性分析

郭帅^{1,2}, 程光^{1,2,3,*}

(1. 东南大学 网络空间安全学院, 南京 211189; 2. 计算机网络和信息集成教育部重点实验室(东南大学), 南京 211189;
3. 网络通信与安全紫金山实验室, 南京 211111)

摘 要: 密码套件是安全传输层协议(TLS)实现安全通信的基石,包含了密钥交换算法、对称密码算法和消息摘要算法,其中对称密码算法被用于实际通信的数据加密。通过对真实流量的采集与分析,得出了不同 TLS 密码套件在现网中的分布情况。设计了一种基于密文图像重构、美国国家标准与技术研究院随机性测试套件、卷积神经网络(CNN)等手段的分析方法,对现网主流对称密码算法(AES、ChaCha20)与其他常见对称密码算法(DES、3DES、RC2、RC4)的密文随机性进行分析。实验结果表明:参与对比的所有对称密码算法在电子密码本(ECB)模式下其密文均具有较差的随机性,无法通过大多数测试;AES 与 ChaCha20 二种主流 TLS 对称密码算法在除 ECB 模式下其密文均具有良好的随机性,对基于 CNN 与随机森林的密码算法识别也具有抵抗能力。研究成果可为 TLS 密码套件的选择与加密流量的深层分析提供参考。

关 键 词: 随机性度量; 密码算法识别; 机器学习; 加密流量识别; 网络测量

中图分类号: TP393

文献标志码: A **文章编号:** 1001-5965(2022)02-0291-10

加密流量的检测与识别是网络测量领域重要而又富有挑战性的研究方向。加密流量在现有互联网总体流量中的占比大幅增加,其中充斥着大量恶意程序流量与网络攻击流量,为网络监管与网络空间治理带来困难。由于密码算法将网络流量的原始信息掩盖,基于明文内容特征的识别方法在面对这些加密流量时难以奏效,加密流量的识别具有了较紧迫的需求与现实意义。加密流量识别在一般意义上指对经过密码算法加密后的数据在网络传输中产生的字节序列进行检测与识别,而被识别的对象则包括加密流量的内容、应用程序、密码算法等。根据加密流量的具体类型与应用场景的不同,又有针对 VPN、安全传输层协议(transport layer security, TLS)及自定义或未知

加密流量的识别。TLS 协议作为现代 Web 网络的主流加密通信协议,被大规模使用,针对 TLS 流量的识别研究在加密流量识别领域具有重要地位,本文通过 TLS 不同加密套件的流量数据随机性分析,对 TLS 加密流量的密码算法识别进行了研究。TLS 密码算法识别本质上属于密码体制识别,指在未知明文、具体密码算法与加密密钥的情况下对密码算法的体制与类型进行识别,是一种被动的唯密文攻击手段。密码体制与算法识别为加密流量的深层次分析提供了另一个角度与方向。

目前,网络流量的密码体制与算法识别主要有数学统计模型与机器学习 2 类,均以不同密码体制或算法产生的密文在字节分布特征上具有可

收稿日期: 2020-08-06; 录用日期: 2020-09-05; 网络出版时间: 2020-12-14 10:06
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201211.1532.001.html
基金项目: 国家重点研发计划(2018YFB1800602); 教育部-中国移动科研基金(MCM20180506); 赛尔网络下一代互联网技术创新项目(NGIICS20190101,NGI20170406)
* 通信作者. E-mail: chengguang@seu.edu.cn

分的差异性为前提,通过提取密文的字节分布特征输入预先建立的统计学模型或神经网络进行识别。密文字节分布的随机性是衡量密码体制与算法性能的重要指标,同样也是对其进行识别的重要特征。2015 年,吴杨等^[1]设计了一种基于码元频数检测、块内频数检测及游程检测的算法对 AES、Camellia、DES、3DES 及 SMS4 密文的随机性度量值取值个数进行了统计分析。2016 年,丁伟和谈程^[2]介绍了一种基于密文分析的密码识别方法,该方法使用支持向量机(SVM)对密码体制进行识别。2018 年,黄良韬等^[3]提出了一种基于随机森林的分层密码体制识别定义系统,相比单层识别方案有约 20% 的提升。2018 年,李洪超^[4]提出了一种基于 SVM、随机森林和决策树的集成学习方法,通过了 15 种随机性测度对电子密码本(ECB)和密文分组链接(CBC)二种加密模式的识别。2019 年,赵志诚等^[5]提出了一种基于随机森林的密码体制识别方法,以美国国家标准与技术研究院(NIST)的随机性测试套件为基础,构建了 54 种密文特征,以显著高于随机识别的准确率完成对 AES、DES、3DES、IDEA、Blowfish 及 Camellia 六种密码体制的两两区分问题。2021 年,王旭等^[6]结合 Relief 特征选择算法和异质集成学习,提出了一种可适应多种密码体制识别情景的动态特征识别方案,比较了 3 类不同密码体制识别情景。

以上工作均以密文字节序列的分布随机性作为基础构建差异特征模型,但大多仅对 DES、3DES 等老旧密码算法与 ECB、CBC 等简单加密工作模式进行分类,对于 TLS 协议中常用的 AES、ChaCha20 密码算法及 GCM 加密模式均没有进行研究与分析,因此,上述工作在加密流量的密码算法识别上作用有限。本文将从实际现网加密流量中的密码算法使用比例出发,分析主流密码算法与其他传统密码算法在密文随机性分布上的特征,并尝试使用 2 种机器学习方法对 AES、RC4 与 ChaCha20 在不同工作模式下的密文进行分类识别,以探究加密流量中密码套件的选择与密码算法在纯密文、有无特征情况下的可识别性问题。

1 TLS 密码套件分析

TLS 密码套件是指在 TLS 连接握手、加密通信与数据完整性检验等过程中被使用的算法集合,包含密钥交换算法、对称密码算法和消息摘要算法等元素^[7]。

1.1 TLS 密码套件组成

根据 TLS 版本的不同,其密码套件组成也不

同,一个典型的 TLS1.2 版本密码套件格式如图 1(a)所示,一个典型的 TLS1.3 版本密码套件组成如图 1(b)所示。

需要说明的是,包含“WITH”字段的均为 TLS1.2 及以下版本支持的密码套件,TLS1.3 版本默认使用 ECDHE 作为密钥交换算法,因此密码套件不需要表明,TLS1.3 版本支持的密码套件编号为 4 865~4 869。虽然 TLS1.2 版本与 TLS1.3 版本加密套件都使用相同的编号格式,但是二者并不通用。经过实际调研,TLS1.3 版本的加密套件只可能是编号为 4 865、4 866、4 867 的 3 种情况。

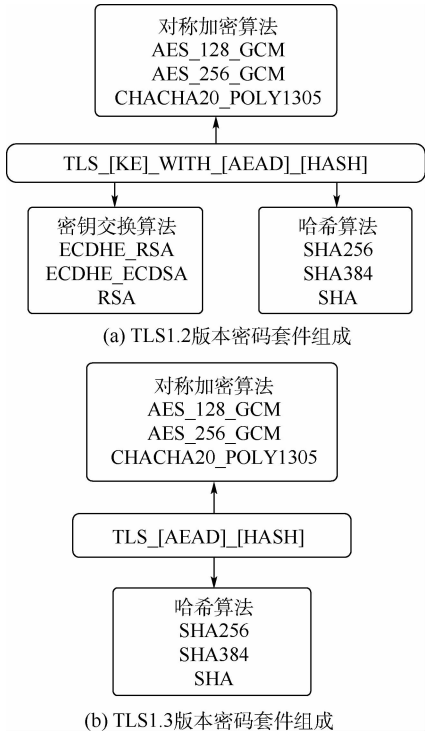


图 1 TLS 密码套件组成

Fig. 1 Composition of TLS cipher suite

1.2 现网流量中的各加密套件占比

本文中使用的现网流量采集自 CERNET 华东北网络中心边界路由,采集周期为 7 d,抽取每天固定时间段内的流量作为样本,经解析得到 TLS 流 1 415 375 条,可提取出加密套件的流共 613 922 条,占比排名前 20 的密码套件如表 1 所示。

分析结果显示,大部分 TLS 流采用的密码套件是 TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256,占比为 59.39%;排名前 5 的密码套件占总体的 88.54%,其中前 4 名均为 AES 的 GCM 模式;在占比前 20 名的密码算法中,除了 1 个 RC4 与 2 个 ChaCha20 算法外,均为 AES 密码算法;这 20 个加密套件占总体样本的 99.73%。除此以外,在这批样本中还观测到了少量的 TLS1.3 版本专用加密套件,说明样本中包含少量 TLS1.3 流量(少于 0.02%)。

表 1 现网流量中各 TLS 密码套件比例

Table 1 Proportion of TLS cipher suites in actual network traffic

密码套件	流数/条	占比/%
TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256	364 600	59.39
TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384	108 966	17.75
TLS_RSA_WITH_AES_256_GCM_SHA384	27 888	4.54
TLS_ECDHE_ECDSA_WITH_AES_128_GCM_SHA256	27 043	4.40
TLS_ECDHE_RSA_WITH_CHACHA20_POLY1305_SHA256	15 080	2.46
TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA384	13 258	2.16
TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA	11 933	1.94
TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA	11 106	1.81
TLS_RSA_WITH_AES_128_GCM_SHA256	9 788	1.59
TLS_RSA_WITH_AES_256_CBC_SHA	4 964	0.81
TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256	3 776	0.62
TLS_ECDHE_ECDSA_WITH_AES_256_CBC_SHA	2 393	0.39
TLS_ECDHE_ECDSA_WITH_AES_256_GCM_SHA384	2 116	0.34
TLS_RSA_WITH_AES_128_CBC_SHA	2 112	0.34
TLS_ECDHE_ECDSA_WITH_AES_128_CBC_SHA	1 721	0.28
TLS_RSA_WITH_AES_256_CBC_SHA256	1 569	0.26
TLS_ECDHE_RSA_WITH_RC4_128_SHA	1 354	0.22
TLS_DHE_RSA_WITH_AES_128_GCM_SHA256	812	0.13
TLS_ECDHE_ECDSA_WITH_CHACHA20_POLY1305_SHA256	688	0.11
TLS_DHE_RSA_WITH_AES_256_GCM_SHA384	668	0.11

上述结果证明,在真实网络的 TLS 加密流量中,使用 AES 作为对称密码算法的情况占绝大部分(97.03%),其余的算法为 ChaCha20 与 RC4。另外,在所有加密套件中均没有发现使用 ECB 模式的密码算法。

2 密码算法的密文随机性检测

密文随机性指经密码算法加密产生的数据在字节分布上的随机程度,密文随机性检测即指对密文字节分布的随机程度进行度量。本文使用了 2 种方法对密文的随机性进行检测:①密文图像重构法是一种主观的随机性检测方法,从宏观角度观察密文的字节分布特征;②基于 SP800-22 测试套件的检测方法则从客观和微观的角度对密文的随机性进行刻画。

2.1 基于密文图像重构法的随机性检测

本节将使用密文图像重构后的可辨别性作为标准对表 2 中所示的密码算法与工作模式进行密文随机性的检测。基于密文图像重构的随机性检测机制是直观且主观的。由于对称密码算法在加密后不会改变原始数据的长度,将图像数据加密后再按照原始尺寸与分辨率重新排列密文比特序列,可将密文重构成与原始数据相同大小的图像。

通过观察密文图像与原始图像,可以直观地感受到不同密码算法与工作模式对数据的影响,也可以观察到不同密码算法在不同的工作模式

表 2 参与实验的密码算法与工作模式

Table 2 Cipher algorithms and working modes involved in experiment

密码算法	工作模式
AES	ECB、CBC、CTR、GCM
DES	ECB、CBC
3DES	ECB、CBC
RC2	ECB、CBC
RC4	
ChaCha20	

下,其密文随机性在宏观上的表现。

2.1.1 实验方法

本节实验采用如图 2 所示的卡通猴图像作为样本,该图像为二值图(只包含黑色和白色 2 种颜色),本质上是由 0(白色)和 1(黑色)组成的比特序列,图像大小为 4 096 × 4 096,共 16 777 216 bit (2 MByte)。图像的比特序列将作为密码算法的输入。

由于表 2 中各密码算法及工作模式的分块大小均为 2 的整倍数,输入样本的数据量可被完整分块无需填充,加密后的数据量和尺寸不变。将密文的比特序列重新按照 4 096 × 4 096 的分辨率进行排列,可将密文组成一幅与原始图像大小相同的图像。通过肉眼观察、对比密文图像与原始图像在内容特征上的相似性,可以得到密文在字节分布上的宏观特征。

2.1.2 实验结果与分析

实验结果如图 3 所示。



图 2 用于实验的图像
Fig. 2 Images for experiments

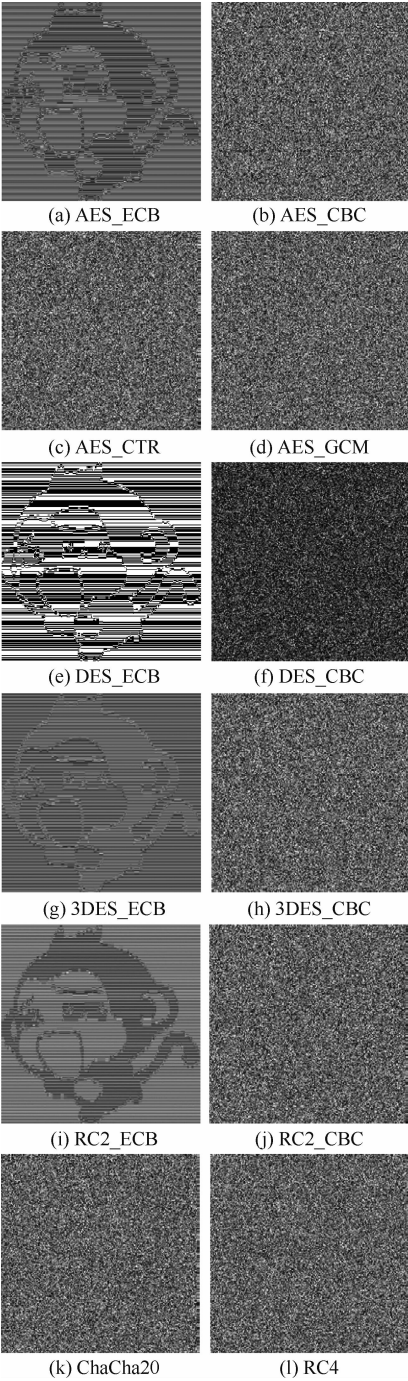


图 3 密文图像实验结果

Fig. 3 Experimental results of ciphertext image

通过实验结果可以看到,对于所有可选分组模式的密码算法,在 ECB 模式下均可观察到原始图像的大致轮廓,而其他模式或密码算法的结果在宏观上均为分布均匀的黑白像素,没有肉眼可以观察到的显著特征。通过分析,本文认为这一现象是由 ECB 本身的工作原理造成的,在 ECB 模式下,明文块加密后按顺序直接拼接。因此,当图像分辨率足够高时(明文长度足够长),明文分块过小(128 bit 或 256 bit)导致仅块内加密并不会影响整体的结构,而其他工作模式或密码算法对并不是分块加密后的明文进行线性地简单排列,破坏了整体结构,使密文无法显现出原始数据具有的轮廓特征。

为了验证密文图像不存在肉眼难以辨认的微小规律性特征,本节将原始图像、AES_ECB、AES_CBC、AES_GCM 与 ChaCha20 密文图像的二维离散傅里叶频率图^[8]作为补充实验,结果如图 4 所示。

通过观察实验结果可发现,原始图像和 ECB 模式下的图像在经过傅里叶变换后,仍旧保留了大量可分辨的高频和低频特征,能量分布并不均匀,证明其密文图像具有较显著的规则结构;其他加密模式的傅里叶变换功率图已无明显的特征,高低频分布均匀,能量分布均匀,证明密文图像无显著的规则结构。

根据上述结果,本文认为参与实验的密码算法在 ECB 工作模式下,其密文的随机性分布较差,具有显著的、肉眼可观测的特征。由于 ECB 工作模式加密时的线性特点,当原始数据的数据

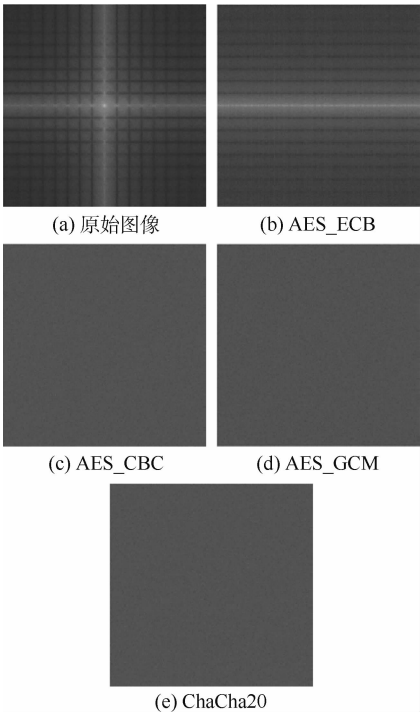


图 4 密文图像的傅里叶频率图

Fig. 4 Fourier frequency diagram of ciphertext image

量足够大时,使用 ECB 工作模式进行加密,密文数据将会一定程度保留原始数据的结构特征,为密文攻击留下隐患。

2.2 基于 SP800-22 测试套件的随机性检测

工业界具有较广泛影响的随机性测试与度量方法是 NIST 于 2010 年修订的 SP800-22 Rev1a 标准^[9],该标准包含 15 个测度指标用于对密码算法和伪随机数生成算法产生数据的随机性进行衡量。SP800-22 Rev1a 标准包含的 15 个指标与测量方法如下:

- 1) monobit test。即整个二进制序列中 0 bit 和 1 bit 的比例,完全随机序列中二者应具有相同或相近的频次。
- 2) frequency within block test。本质上还是 monobit test,只不过该测度将待测数据分块,再统计并在每块内执行 monobit test。
- 3) runs test。 n 个相同的二进制位被视为 1 个长度为 n 的游程。游程检测通过统计一段密文中所包含的不同长度游程的数量,判断其是否与理想随机数据的游程数量一致。
- 4) longest run ones in a block test。分块内最长连续 1 bit 序列的长度。
- 5) binary matrix rank test。将数据的比特序列组成矩阵后,计算矩阵的秩。
- 6) dft test。将数据通过傅里叶变换转换到频域,检测其频率谱中能量尖峰的分布情况,目的是检测待检验信号的周期性。
- 7) non overlapping template matching test。用来测试数据中是否出现了过多的非周期模式序列。对于 n bit 长度的模式序列,使用 1 个 n bit 长度的滑动窗口进行匹配,如果在当前位置匹配失败,窗口向后滑动 1 bit,如果当前位置匹配一致,窗

- 口向后滑动 n bit,继续进行匹配。
 - 8) overlapping template matching test。该指标与 non overlapping template matching test 相似,用来测试数据中是否出现了过多的非周期模式序列,不同之处在于该指标在匹配成功后只滑动 1 bit。
 - 9) linear complexity test。计算输入数据的线性复杂度,评估其与完全随机序列的差距。
 - 10) serial test。该指标检测所有连续 m bit 的序列在输入数据中出现的频次。
 - 11) approximate entropy test。计算输入数据的近似熵,评估其与完全随机序列的差距。
 - 12) cumulative sums test。检测输入数据中的部分比特和是否太大或太小。
 - 13) random excursion test。检测输入数据中某一状态出现的次数是否远超完全随机序列。
 - 14) random excursion variant test。检测输入数据中某一状态在 1 次随机游程中出现的次数是否远超完全随机序列。
 - 15) maulers universal test。检测输入数据是否可以被压缩算法显著压缩。
- SP800-22 Rev1a 标准采用显著性检验模型对输入数据的随机性进行度量,即检验输入数据与假设的完美随机序列之间的统计学差距。由于假设检验不属于本文主要内容,在此不再赘述。当某一指标的指数大于 0.01 时,有 99% 的置信度认为输入数据在该指标上符合完全随机性,指数越接近 1,说明输入数据在该指标上越接近完全随机序列。
- 基于 SP800-22 Rev1a 测试套件的随机性检测将图 2 的原始数据经过 AES_GCM、AES_ECB、RC4 和 ChaCha20 密码算法加密后,输入 SP800-22 Rev1a 随机性测试套件,对密文的 15 个随机性指标进行检测,其结果如表 3 所示。通过实验结果

表 3 图 2 在 SP800-22 测试套件上的随机性测度检测结果
Table 3 SP800-22 randomness measurement test results of Fig. 2

指标	原始数据		AES_ECB		ChaCha20		AES_GCM		RC4	
	指标	结论	指数	结论	指数	结论	指数	结论	指数	结论
monobit test	0	不通过	0	不通过	0.297	通过	0.403	通过	0.827	通过
frequency within block test	0	不通过	0	不通过	0.603	通过	0.181	通过	0.105	通过
runs test	0	不通过	0	不通过	0.764	通过	0.504	通过	0.926	通过
longest run ones in a block test	0.002	不通过	0.003	不通过	0.572	通过	0.611	通过	0.691	通过
binary matrix rank test	0	不通过	0	不通过	0.404	通过	0.787	通过	0.489	通过
dft test	0	不通过	0	不通过	0.013	通过	0.011	通过	0.829	通过
non overlapping template matching test	0.001	不通过	0.970	通过	1.000	通过	1.000	通过	0.983	通过
overlapping template matching test	0	不通过	0	不通过	0.887	通过	0.370	通过	0.567	通过
linear complexity test	0	不通过	0	不通过	0.303	通过	0.939	通过	0.047	通过
serial test	0	不通过	0	不通过	0.055	通过	0.257	通过	0.888	通过
approximate entropy test	0	不通过	0	不通过	0.054	通过	0.257	通过	0.887	通过
cumulative sums test	0	不通过	0	不通过	0.413	通过	0.290	通过	0.696	通过
random excursion test	0.001	不通过	0.173	通过	0.047	通过	0.057	通过	0.325	通过
random excursion variant test	0.414	通过	0.160	通过	0.187	通过	0.042	通过	0.035	通过
maulers universal test	0	不通过	0	不通过	0.335	通过	0.011	通过	0.048	通过

可知, AES_GCM、RC4 及 ChaCha20 这 3 种主流 TLS 密码算法在随机性上表现优秀, 通过了 SP800-22 Rev1a 测试套件的全部指标, 而 AES_ECB 模式仅通过了 3 项, 其余指标均未通过, 与原始数据几乎相同。因此, ECB 模式无法完全掩盖原始数据具有的非随机特征, 与其他密码算法相比在密文的随机性上表现较差, 具有可区分性, 容易遭受密码分析攻击。

3 密码算法类型识别的可行性分析

应用于密码算法识别与加密流量分类的主要机器学习方法有 2 类: 一类是以 SVM^[10]、决策树^[11]、随机森林^[12] 等为代表的传统机器学习方法; 另一类是以多层感知机 (MLP)^[13]、卷积神经网络 (CNN)^[14]、循环神经网络 (RNN)^[15] 等模型为代表的深度学习方法。使用传统的机器学习方法需要通过特征工程提取原始数据的相关特征作为训练数据, 而使用深度学习方法可以直接输入原始数据, 依托深度学习模型的自主特征挖掘能力进行训练。本文将分别使用 CNN 和随机森林这 2 种在相关问题中被证明有效的模型, 进行基于特征和原始数据的无特征密码算法类型与工作模式的识别可行性分析。

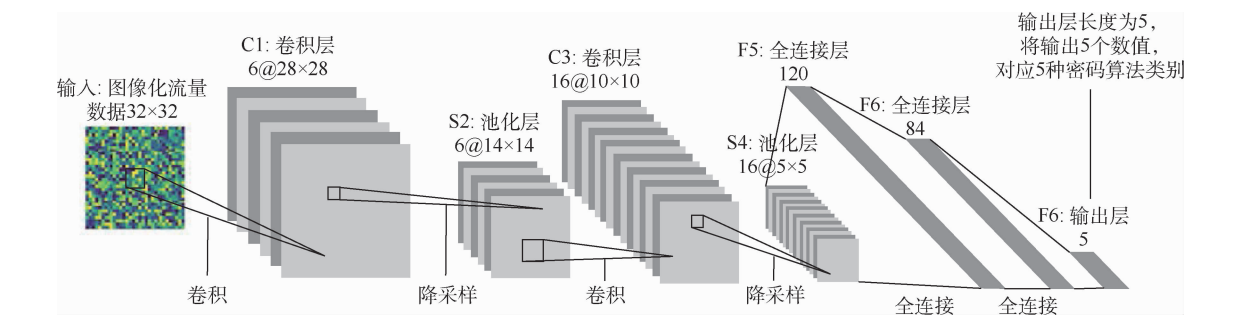


图 5 用于密码算法识别的 CNN 结构
Fig. 5 CNN structure for cipher algorithm recognition

3.1.2 基于随机森林的有特征识别

随机森林是一种集成学习分类方法^[16], 通过建立多个独立模型的组合来解决一个分类问题, 其会生成多个分类器, 每个分类器各自学习和做出决策, 这些决策最后会通过投票形成一个单一的分类结果, 因此在一般情况下, 集成学习的结果将优于一个单独分类器。随机森林使用多个独立决策树的决策来确定最后的分类结果, 因此, 其也是一种集成学习方法。

本文将使用决策树作为分类模型, 进行不同密码算法与工作模式的识别。随机森林是基于特征的分类器模型, 不能直接输入原始数据, 需要进行特征的提取。本文将使用 SP800-22 Rev1a 的全

3.1 识别方案

3.1.1 基于 CNN 的无特征识别

CNN 是包含卷积计算且具有深度结构的前馈神经网络, 在计算机视觉、图像分析和处理等领域发挥了巨大作用, 其接收 1 个多通道矩阵作为输入, 经过卷积层、池化层、全连接层等网络结构, 最后输出 1 个向量作为分类结果。LeNet5^[14] 是最经典的 CNN 网络模型, 最早被应用于手写数字的识别, 是一个十分类问题, 其输出层尺寸为 10×1 。为适应实验需求, 本文对其输出层结构进行了修改, 具体结构组成依次为输入层 (32×32 , 1 通道)、二维卷积层 (28×28 , 6 通道)、最大池化层 (14×14 , 6 通道)、二维卷积层 (10×10 , 16 通道)、最大池化层 (5×5 , 6 通道)、全连接层 (120×1 , 单通道)、全连接层 (84×1 , 单通道) 及输出层 (5×1 , 单通道), 详细结构如图 5 所示。

该模型对应一个五分类问题, 即对 AES_ECB、AES_CBC、AES_GCM、RC4、ChaCha20 五种不同的密码算法与模式进行识别。模型的输入为原始流量数据经过处理后生成的尺寸为 32×32 的字节矩阵, 输出为 1×5 的向量, 代表最终的密码算法类型分类结果。整个训练过程共进行 200 次迭代。

部测度作为样本的特征, 组成一维特征向量 $\mathbf{F}$:
$$\mathbf{F} = (x_1, x_2, x_3, \dots, x_n)$$
 (1)
式中: x_i 代表 SP800-22 Rev1a 测试套件中的 1 个测度值; n 为特征向量长度, 本文中 $n = 15$ 。

使用样本特征向量组成各密码算法的样本数据集 S_{algo} , 样本数据集包含了该密码算法或工作模式用于训练和测试的所有样本:

$$S_{\text{algo}} = (F_1, F_2, F_3, \dots, F_k)$$
 (2)
式中: algo 指某一种密码算法或工作模式; F_j 为第 j 个样本的特征向量; k 为样本数量, 在本文中 $k = 5\,000$, 即每一类密码算法或工作模式有 5 000 个样本。

为了完成随机森林的二分类任务, 需要从各

密码算法或工作模式数据集中抽取 2 个组成用于模型训练和测试的扩展数据集 T 与标签集 L :

$$T = (S_{\text{algo}}^1, S_{\text{algo}}^2) \tag{3}$$

式中: S_{algo}^1 与 S_{algo}^2 分别代表一种密码算法或工作模式的样本集。

$$L = (A_1, A_2, A_3, \cdots, A_m) \quad A \in \{1, 2\} \tag{4}$$

式中: A_m 代表扩展数据集 T 中第 m 个样本的密码算法或工作模式类型,取值为 1 或 2,代表二分类问题中的第 1 类或第 2 类。

基于随机森林的有特征识别采用了十折交叉验证方法,因此,只需要对扩展数据集 T 进行随机打乱排序,而不需要手动进行训练集和测试集的划分。在进行模型的训练时,输入打乱顺序后的扩展数据集 T ,模型输出预测标签集 L_{pre} ,将 L_{pre} 与原始标签集 L 对比可得到模型的分类准确率。

3.2 实验方法

3.2.1 数据采集方法

本节实验采用 TLS1.2 加密网络流量作为基础数据样本,根据第 1 节对现网流量中密码套件的占比分析,实验采用如表 4 所示的 4 种密码套件产生流量。此外,为了验证模型的有效性,除了表 4 中列出的 4 种密码套件,本文还加入了 AES_ECB 模式,即总共 5 种密码算法。

需要说明的是,在 TLS 密码套件中并不存在使用 ECB 的选项,因此,无法通过模拟产生 TLS 流量来获取 AES_ECB 加密的流量。本文将使用 OpenSSL 中的 AES_ECB 加密模块对原始数据进行加密,直接获得 ECB 工作模式下的样本。这种方法在本质上与 TLS 产生加密数据的过程类似,可以避免因样本产生方法不同而对实验结果产生影响。

4 种密码套件的主要密码算法分别是 AES_CBC、AES_GCM、ChaCha20 及 RC4,使用比例在本文采集的现网流量中达到 90%。因此,这 4 种密码算法具有足够的代表性,如果能够对其进行区分,具有较大的现实意义。TLS 流量由本地模拟环境产生,原始内容为 HTTP 形式的响应数据,其身体部分为长度 1 000 ~ 1 200 Byte 的邮件文本序列^[17],加上约 160 Byte 的 HTTP 头部,整个负载

加密后长度在 1 300 ~ 1 500 Byte 以内。通过反向代理,将 HTTP 流量封装成 TLS 流量,本节实验仅使用 1 024 Byte 长度的加密数据,因此,1 个样本仅需 1 次 HTTP 请求即可。AES_ECB 使用同样的原始数据,直接通过 OpenSSL 的加密模块生成 1 024 Byte 长度的加密数据。

本节实验一共使用 20 000 个密文样本,5 种密码算法或工作模式的样本数量相同(每种 5 000 个样本)。在基于 CNN 的无特征识别中,从整个数据集中随机抽取 80% 作为训练样本,20% 作为测试样本;在基于随机森林的有特征识别中,从整个数据集中提取 2 种密码算法或工作模式的样本构成扩展数据集,由于使用十折交叉验证,无需手动划分训练集和测试集。

3.2.2 数据预处理方法

按照图 6 所示的方法,对 TLS 样本流量的应用负载部分进行提取。首先,根据各密码算法的加密规则去除其头部和尾部的初始化向量、MAC 校验码等附加部分,剥离出纯密文部分。然后,将纯密文部分裁剪为 1 024 Byte 长度(由于流量生成时长度最低为 1 300 Byte,不存在密文长度不够时的填充问题,避免干扰),AES_ECB 的密文无需处理,可直接生成 1 024 Byte 长度的样本。

在基于 CNN 的无特征识别中,将 1 024 Byte 的样本数据按顺序排列成一副尺寸为 32 像素 × 32 像素的 256 色阶的单通道灰度图(即每个像素为 1 Byte 大小,代表 0 ~ 255 的不同黑白色阶,32 × 32 = 1 024)。按照分组密码算法 16 Byte 的分组长度,该图片的 1 行包含 2 个密文分组,每张图包含 64 个密文分组,且原始密文数据的顺序按从左到右、从上到下进行排列。

图 7 为一个经过图像化处理的样本,包含 1 024 Byte 的纯密文。

在基于随机森林的有特征识别中,将 1 024 Byte

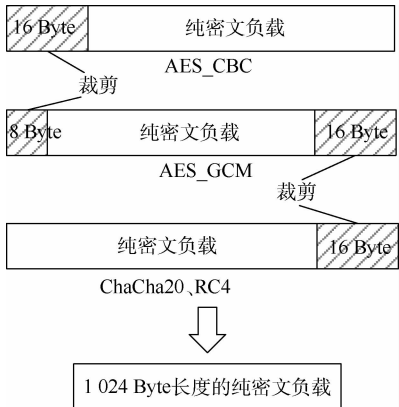


图 6 样本预处理方法

Fig. 6 Sample pretreatment method

表 4 实验采用的密码套件	
Table 4 Cipher suite used in experiment	
编号	密码套件名称
0xc014	TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA
0xc030	TLS_ECDHE_RSA_WITH_AES_256_GCM_SHA384
0xcc8	TLS_ECDHE_RSA_WITH_CHACHA20_POLY1305_SHA256
0xc011	TLS_ECDHE_RSA_WITH_RC4_128_SHA

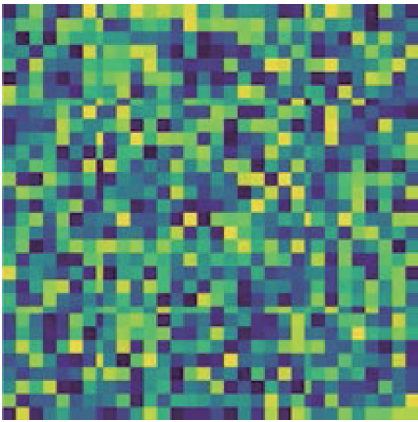


图 7 经处理后的样本
Fig. 7 Processed samples

长度的样本数据输入 SP800-22 Rev1a 测试套件, 获取每一个样本的 15 个随机测度, 并按照式(1) ~ 式(4)生成对应的扩展数据集 T 和标签集 L , 作为随机森林分类器的输入。

3.3 实验结果

3.3.1 CNN 模型识别结果

图 8 为训练过程中训练集准确率和测试集准确率的变化。可以看到, 随着迭代次数的增加, 训练集准确率持续上升, 在第 125 个迭代时逼近 100%, 但测试集准确率却始终徘徊在约 20% 处, 且整体呈现下降趋势, 这说明在训练的早期模型可能就已经出现了过拟合现象。出现过拟合现象的原因有很多, 如模型过于简单、样本过少、训练时间过长、样本噪声过大等。由于本节实验训练过程中, 训练集准确率持续上升, 测试集准确率一直处于下降状态, 说明模型可以挖掘和学习到训练集样本的特征, 但是无法将这种特征推广到测试集或其他数据集, 也说明通过深度学习模型从纯密文样本挖掘到的全局特征是“随机”的, 即对于各种密码算法, 其密文特征并没有统一模式, 这与密码算法的本质要求是一致的。

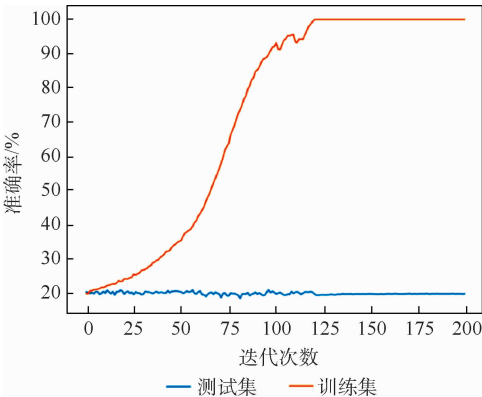


图 8 训练过程中准确率的变化

Fig. 8 Change of accuracy during training

最终, 该模型在 AES_ECB、AES_CBC、AES_GCM、ChaCha20、RC4 这 5 种密码算法识别问题上的准确率为 20.92%, 分类结果的混淆矩阵如图 9 所示, 该模型在此处的五分类问题上与随机猜测的准确率一致。因此, 本文认为 LeNet5 这一深度学习模型在本文的实验环境下, 无法解决 AES_ECB、AES_GCM、AES_CBC、ChaCha20 及 RC4 纯密文数据的五分类问题, 即使是随机性表现较差的 AES_ECB, 也无法通过纯密文负载与其他密码算法进行区分。

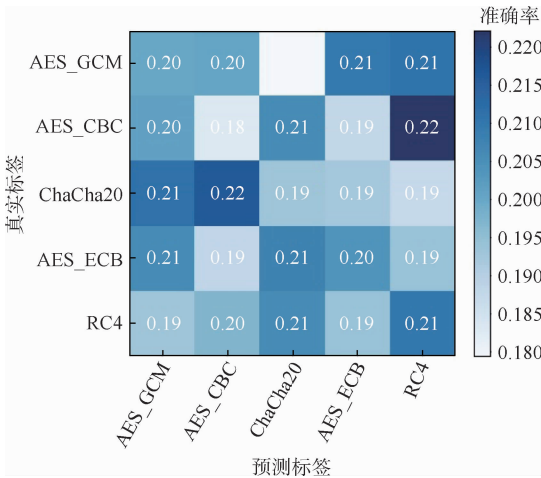


图 9 分类结果混淆矩阵

Fig. 9 Confusion matrix of classification results

3.3.2 随机森林模型识别结果

本文使用随机森林进行密码算法间的二分类问题, 共有 5 种密码算法需要分类, 因此, 可以形成 10 个不同的二分类问题, 即随机森林的输入为 10 个扩展数据集 T 和标签集 L 。随机森林的二分类识别结果如表 5 所示。

通过表 5 中的结果可知, 除 AES_ECB 外, 其余密码算法或工作模式的二分类准确率均约为 50%, 与随机猜测的准确率几乎相同。而包含 AES_ECB 的二分类准确率均高于 50%, 但均低于 60%, 识别效果略高于随机猜测。

表 5 随机森林二分类准确率

Table 5 Accuracy of random forest classification		
类别 1	类别 2	准确率/%
AES_ECB	AES_CBC	55.3
	AES_GCM	58.2
	ChaCha20	57.1
	RC4	57.9
AES_CBC	AES_GCM	48.3
	ChaCha20	47.9
AES_GCM	RC4	50.0
	ChaCha20	48.7
ChaCha20	RC4	50.1
	RC4	48.4

值得注意的是,该结果所表现出的 AES_ECB 与其他密码算法的随机性差异并没有表 3 中显著。本文认为这与原始数据本身的随机性有关,表 3 中的结果由图 2 经各密码算法加密后进行随机性检测得出,图 2 的原始数据在绝大多数随机性测度上都接近 0。同时, AES_ECB 无法掩盖原始数据的字节分布特征,导致经 AES_ECB 加密后的密文与其他密码算法的密文存在巨大差异。在本节实验中,原始数据来源于网络邮件数据集,在网络流量背景下比图 2 更具有代表性,其本身的随机性就比图 2 强,因此,经过 AES_ECB 加密后与其他密码算法的差异没有表 3 中显著。虽然对 AES_ECB 与其他密码算法的分类结果不显著,但仍高于随机猜测,证明本文模型在密码算法的分类上具有一定有效性;而对于除 AES_ECB 外的其他密码算法或工作模式,在本文中的实验环境下无法做到有效的区分。

4 结 论

本文通过对现网流量中 TLS 密码套件使用情况,以及通过实验对各类对称密码算法在不同工作模式下的密文随机性及抵抗密码体制识别能力上的表现进行比较,得出以下结论:

1) 在现网流量中, AES 是 TLS 协议中使用最广泛的对称密码算法, 占总体的 97.03%; GCM 是 AES 算法中最常使用的工作模式, 占总体的 90.98%。现有关于密码算法识别的工作大多没有涉及 AES_GCM 与 ChaCha20 这 2 种主流密码算法, 因此, 并不适用于 TLS 网络流量中的密码算法别问题。

2) TLS 中各类对称密码算法在 ECB 工作模式下均具有较差的密文随机性, 其密文带有显著的原始数据结构特征。在非 ECB 模式下, 密文随机性表现良好, 密文的傅里叶频率图能量分布均匀, 其中 AES_GCM、AES_CBC、ChaCha20 及 RC4 密文均可通过 SP800-22 Rev1a 的全部指标。

3) AES_ECB、AES_CBC、AES_GCM、ChaCha20 及 RC4 这 5 种密码算法或其分组模式均可在一定程度上抵御基于无特征识别的分析攻击, 但 AES_ECB 模式无法抵御基于有特征识别的分析攻击。

4) 进行加密流量的分析时, 应同时考虑数据的非加密与加密部分、综合数据的低熵区与高熵区特征。

5) 在进行 TLS 加密通信时, 应优先选择带有 AES_GCM 或 ChaCha20 对称密码算法的密码套

件, 以对抗基于随机性特征的密码体制识别攻击。

参考文献 (References)

[1] 吴杨,王韬,邢萌,等. 基于密文随机性度量值分布特征的分组密码算法识别方案[J]. 通信学报, 2015, 36 (4) : 150-159.

WU Y, WANG T, XING M, et al. Block ciphers identification scheme based on the distribution character of randomness test values of ciphertext[J]. Journal on Communications, 2015, 36 (4) : 150-159 (in Chinese).

[2] 丁伟,谈程. 一种基于密文分析的密码识别技术[J]. 通信技术, 2016, 49 (10) : 1382-1386.

DING W, TAN C. An approach of identifying cipher based on ciphertext analysis[J]. Communications Technology, 2016, 49 (10) : 1382-1386 (in Chinese).

[3] 黄良韬,赵志诚,赵亚群. 基于随机森林的密码体制分层识别方案[J]. 计算机学报, 2018, 41 (2) : 382-399.

HUANG L T, ZHAO Z C, ZHAO Y Q. A two-stage cryptosystem recognition scheme based on random forest[J]. Journal of Computers, 2018, 41 (2) : 382-399 (in Chinese).

[4] 李洪超. 基于密文特征的密码算法识别研究[D]. 西安: 西安电子科技大学, 2018.

LI H C. Cipher-text features based cipher system recognition [D]. Xi'an: Xidian University, 2018 (in Chinese).

[5] 赵志诚,赵亚群,刘凤梅. 基于随机性测试的分组密码体制识别方案[J]. 密码学报, 2019, 6 (2) : 177-190.

ZHAO Z C, ZHAO Y Q, LIU F M. Scheme of block ciphers recognition based on randomness test[J]. Journal of Cryptologic Research, 2019, 6 (2) : 177-190 (in Chinese).

[6] 王旭,陈永乐,王庆生,等. 结合特征选择与集成学习的密码体制识别方案[J]. 计算机工程, 2021, 47 (1) : 139-145.

WANG X, CHEN Y L, WANG Q S, et al. Cryptosystem identification scheme combining feature selection and ensemble learning[J]. Computer Engineering, 2021, 47 (1) : 139-145 (in Chinese).

[7] DIERKS T. The transport layer security (TLS) protocol version 1.2 [EB/OL] (2020-01-21) [2020-07-05]. <https://tools.ietf.org/html/rfc5246>.

[8] BRACEWELL R. The Fourier transform and its applications [J]. American Journal of Physics, 2002, 34 (8) : 712.

[9] RUKHIN A, SOTA J, NECHVATAL J, et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications; SP800-22 Rev1a. 1a [S]. Washington, D. C. : National Institute of Standards and Technology, 2010.

[10] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20 : 273-297.

[11] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1 (1) : 81-106.

[12] HO T K. Random decision forests[C] // Proceedings of 3rd International Conference on Document Analysis and Recognition. Piscataway: IEEE Press, 1995 : 278-282.

[13] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The elements of statistical learning; Data mining, inference and prediction[M]. Berlin: Springer, 2009.

[14] LECUN Y,BOTTOU L,BENGIO Y,et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.

[15] GRAVES A,LIWICKI M,FERNÁNDEZ S,et al. A novel connectionist system for unconstrained handwriting recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2008,31(5):855-868.

[16] ROKACH L. Ensemble-based classifiers[J]. Artificial Intelligence Review,2010,33(1-2):1-39.

[17] WILLIAM W. Enron email dataset[DS/OL]. [2020-08-01]. <https://www.cs.cmu.edu/~w/enron/>.

Randomness of traffic data in TLS cipher suite

GUO Shuai^{1,2}, CHENG Guang^{1,2,3,*}

(1. School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China;

2. Key Laboratory of Computer Network and Information Integration(Southeast University), Ministry of Education, Nanjing 211189, China;

3. Purple Mountain Laboratory for Network and Communication Security, Nanjing 211111, China)

Abstract: Cipher suite is the cornerstone of transport layer security (TLS) to realize secure communication, which includes asymmetric cipher algorithm, symmetric cipher algorithm and message digest algorithm, among which symmetric cipher algorithm is used for data encryption in actual communication. Through the collection and analysis of real traffic, this paper obtains the distribution of different TLS cipher suites in the existing network. Then, an analysis method based on image ciphertext reconstruction, NIST randomness test suite and convolutional neural network (CNN) is designed to analyze the ciphertext randomness of mainstream symmetric cipher algorithms (AES, ChaCha20) and other common symmetric cipher algorithms (DES, 3DES, RC2, RC4). The experimental results show that the ciphertexts of all the symmetric cipher algorithms participating in the comparison have poor randomness in the electronic codebook (ECB) mode and cannot pass most tests. AES and ChaCha20, two mainstream TLS symmetric cipher algorithms, have good randomness in ciphertext except ECB mode, and have resistance to cipher algorithm recognition based on CNN or random forest. Relevant research can provide reference for the deep analysis of TLS cipher suite selection and encrypted traffic.

Keywords: randomness measurement; cipher algorithm recognition; machine learning; encrypted traffic identification; network measurement

Received: 2020-08-06; Accepted: 2020-09-05; Published online: 2020-12-14 10:06

URL: kns.cnki.net/kcms/detail/11.2625.V.20201211.1532.001.html

Foundation items: National Key R & D Program of China (2018YFB1800602); Ministry of Education-China Mobile Research Fund Project (MCM20180506); CERNET Innovation Project (NGHCS20190101, NGH20170406)

* Corresponding author. E-mail: chengguang@seu.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0404

基于情感对象识别和情感规则的 微博倾向性分析

王泽辰¹, 王树鹏^{1,*}, 孙立远², 张磊¹, 王勇¹, 郝冰川¹

(1. 中国科学院信息工程研究所, 北京 100193; 2. 国家计算机网络应急技术处理协调中心, 北京 100085)

摘 要: 微博平台数据中含有大量反映用户情感喜恶的信息,对于涉及博文倾向性分析的应用尤为重要。现有的分析方法往往聚焦在博文情感的简单分类上,无法分析特定类型实体的微博倾向性。为解决微博倾向性分析问题,实现博文立场判定,采用半监督学习的方法,通过协同训练和主动学习,训练实体识别模型,并构建基于主成分分析的情感规则,提取句子的主成分,将口语化的文本规范化为指定格式。再利用指向性实体的正负面性、情感词的褒贬义及情感词充当的句子成分,实现情感分类的更深层次分析——立场判定。针对实际问题进行立场判定实验,在不同规模数据集上的自对比实验和他比实验显示,随着标注实体的博文数量增加,模型对博文立场判断的正确率持续提升,而且所提方法判断博文立场的正确率显著高于对比方法,相较已有研究方法分别提高了 2.79% 和 10.00%。

关 键 词: 情感分析; 立场判定; 半监督学习; 倾向性; 情感规则; 协同训练; 主动学习

中图分类号: P391

文献标志码: A

文章编号: 1001-5965(2022)02-0301-10

微博作为中文社交平台,具有强大的影响力和渗透力。微博用户不断增加,2020 年第 1 季度财报显示,微博日活跃用户达 2.41 亿,月活跃用户达 5.5 亿。这些用户在平台上建立关系、获取信息及生产大量内容。在博文文本信息中,大量博文包含带有明显情感色彩的内容。挖掘这些文本背后用户的观点与倾向,可以判断流行趋势和热点,有助于企业分析消费者购买倾向,进行精准化营销,也使得政府可以对网民的舆论立场变化做出及时反应。因此,对博文文本进行倾向性分析,进而完成对博文立场的判定,是情感分析中一项重要的研究任务。

情感分析以对特定文本进行处理及挖掘其中情感色彩为目的,是自然语言处理的重要分支,近

年来受到了广泛关注。提取情感词是分析微博所表达情感最直接的方法。微博文本具有长度较短、内容形式多样、观点倾向性强、表达方式口语化、普遍缺少上下文信息等特点。针对微博数据有许多情感分析的研究,结合情感词典对微博进行分析的方法较为直观。通过对微博中包含的情感词进行分类,可以简单地概括其中蕴含的情绪。但是 1 条微博所表达的信息不仅与其包含的情感词相关,即使 2 条微博包含完全相同的情感词,当情感词所指向的对象不同时,微博的观点与倾向性会存在明显差异。例如,电子产品制造商可能只希望关注用户与手机、电脑等相关的微博情感信息,而不关心对食品、日用品等的购买倾向;在舆情方面,要分析网民在讨论社会热点问题时的

收稿日期: 2020-08-09; 录用日期: 2020-09-25; 网络出版时间: 2021-03-09 20:12
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210309.0928.002.html
基金项目: 国家自然科学基金(61931019)
* 通信作者. E-mail: wangshupeng@iie.ac.cn

引用格式: 王泽辰, 王树鹏, 孙立远, 等. 基于情感对象识别和情感规则的微博倾向性分析[J]. 北京航空航天大学学报, 2022, 48(2): 301-310. WANG Z C, WANG S P, SUN L Y, et al. Weibo tendency analysis based on sentimental object recognition and sentimental rules [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 301-310 (in Chinese).

立场,不仅需要辨别微博表达的态度是积极还是消极,还需要确定微博支持或反对的实体是什么。另外,中文词汇的多义性使得同一词语在不同语境下指代的实体含义可能不同。随着网络语言的发展,用于实体表示的新词也层出不穷,用户常常用品牌缩写、谐音、流行语等进行指代,这些依赖人工标注的实体往往需要很大的工作量。因此,如何针对海量微博数据进行倾向性分析,从而实现立场判定,是目前亟须解决的问题。

为解决微博倾向性分析的问题,本文将深度学习的方法与情感规则相结合,基于句法规则进行博文情感分析,指定相应的情感规则,结合协同学习和主动学习的方法,可以在仅依赖少量标注数据的前提下准确判断博文对指定类型实体表达的情感。

本文主要有以下 3 点创新:

- 1) 利用半监督学习的方式,通过协同训练加主动学习的方式,同时利用半监督学习结合主动学习的模式,确定指向性实体集。
- 2) 不同于传统的情感规则,本文提出了基于主成分分析的情感规则,通过识别指定方向实体,结合情感词对博文进行倾向性判断。
- 3) 生成指定类别实体集,结合情感规则,判断博文的立场,实现对博文更深层次的分析。

1 相关工作

网络舆情中的情感识别问题得到了学界广泛关注。网络社交平台数据是情感识别任务的重要数据来源。例如,基于 Twitter 平台数据,Giachanou 等^[1]建立了情感分析方法,训练、分析推文内容的情感倾向性。类似地,微博作为受众广、数据规模大的中文平台,吸引了许多学者对其数据开展研究。

微博情感分析方法主要有 3 种:

- 1) 基于语义词典的方法。该方法需要先构建微博情感分析数据库,一般包含多个词典和句法规则库,再利用知识库进行聚合计算。例如,王志涛等^[2]提出了一种基于词典和规则集的中文微博情感分词方法,将词典和语义规则结合进行情感分析。王灿伟^[3]提出了结合情感词典和感情符号来计算微博情感值的方法,实现主题归类和情感分类。语义词典通过构造情感词典,筛选出微博文本中的情感词,计算文本片段的情感权值,从而判断微博的情感倾向。Ebrahimi 等^[4]提出了将情感极性融入到情感对象和情感立场中的方法,通过对数线性联合建模实现立场判定。

- 2) 传统机器学习方法。该方法普遍应用于情感分类,其核心思想是构建特征向量,找到特征与分类结果之间的关联。朴素贝叶斯 (naive Bayes)、最大熵 (maximum entropy, ME)、支持向量机 (support vector machine, SVM) 等方法常用于情感分类^[5]。Pang 等^[6]使用了电影评论作为数据,采用以上 3 种机器学习方法进行情感分类。莫雨洁等^[7]使用 SVM、随机森林和梯度提升决策树对文本特征进行立场检测,并利用特征分类器进行立场融合。这些方法具有训练时间短及特征维度高的特点,但不能充分利用上下文信息和语法信息,且需要大量手工标注。

- 3) 基于深度学习的方法。该方法对微博博文进行分词,将其表示为词向量,通过深度神经网络提取语义信息,构建情感表征向量,并利用微博情感表征完成情感分类任务。深度学习模型在情感分析上取得了很好的效果^[8],特别是基于长短期记忆网络 (long and short term memory, LSTM) 方法,在长语句分析上取得了很好的结果^[9]。除此之外,还有基于双重注意力模型^[10]、监督学习^[11]、集成学习^[12]的方法。

通过对上述已有工作进行总结和分析,可以将基于语义词典、传统机器学习和深度学习的微博情感分析方法的特点进行对比,结果如表 1 所示。由表 1 可知,深度学习方法能够充分利用文本信息,往往具有更好的效果。基于深度学习的情感分析又可细分为无监督学习、有监督学习和半监督学习 3 种方法。

在无监督学习方法中,所有数据都缺少情感标注。Turney^[13]提出了情感倾向性分类的方法,通过选取指定词汇,分别计算每个词与情感词典中积极、消极情感词的交叉熵,以 ME 值作为情感词分类判断标准。除此之外,还有基于语法分析^[14]、句法模式^[15]等方法的研究。上述方法可以发掘文本数据中内在的词汇情感规律,节约人工标注的成本。

表 1 微博情感分析方法的特点

Table 1 Features of Weibo sentiment analysis methods

情感分析方法	特点
语义词典	包含多个词典和句法规则库;利用知识库进行聚合计算;需要构建微博情感分析数据库
传统机器学习	构建特征向量;找到特征与分类结果间的关联;需要大量手工标注
深度学习	分词并将词语表示为词向量;深度神经网络提取语义信息;构建情感表征向量;需要较长训练时间

有监督学习情感分类问题主要关注特征选择和分类器设计。例如,文献[16]中采用了词性特征,Cambria 等^[17]引入了符号特征和词嵌入特征。基于商品评论中产品特征信息的情感分类采用有监督学习方法^[18],具有较高的运行效率。但是,如果在有监督学习中出现分类错误,那么之后的学习都会受到这个错误的影响。为了解决这个问题,有学者提出了半监督学习方法避免准确率的下降。

Sindhvani^[19]和 Liu^[20]等分别提出了采用半监督学习方法进行情感分析。半监督学习包括自训练、协同训练等方法。自训练^[21]方法是较早提出的一种半监督学习方法,先训练初始分类器,再使用该分类器对未标注的数据进行标注,选取分类准确度较高的样本确认标注,直到经过多轮迭代后所有样本都完成标注为止。半监督学习方法仅需要标记少量数据,适用于数据量大的任务。

对基于深度学习的方法进行总结和分析,将基于无监督学习、有监督学习和半监督学习的深度学习情感分析方法的特点进行归纳,结果如表 2 所示。

由表 2 可知,采用半监督学习方法,在情感分析任务上可以取得较好的效果。基于情感分析研究现状,本文提出了基于情感对象识别与情感规则的倾向性分析(orientation analysis based on sentiment object recognition and sentiment rules, OASOSR)算法,不仅能够确定博文中的情感词,还可以分析情感词指向的实体是否为目标实体。

表 2 基于深度学习的情感分析方法特点对比
Table 2 Comparison of features of sentiment analysis methods based on deep learning

情感分析方法	特点
无监督学习	发掘文本数据中内在的词汇情感规律;无需人工标注
有监督学习	运行效率较高;出现分类错误时后续训练会受到影响
半监督学习	仅需要标记少量数据,适用于数据量大的任务

2 OASOSR 算法

OASOSR 算法框架如图 1 所示,主要分为 2 个部分:①基于协同训练和主动学习的实体识别模型;②基于情感词典和句法分析的微博倾向性分析。

基于协同训练和主动学习的实体识别模型部分采用半监督学习方法,将少量已标注的微博文

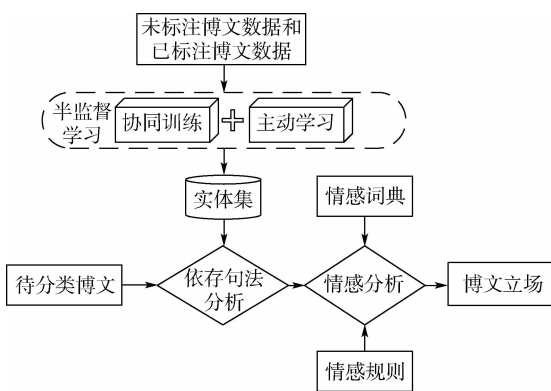


图 1 OASOSR 算法总体架构
Fig. 1 Algorithm architecture of OASOSR

本数据集作为初始输入,协同训练 2 个不同的实体识别模型。为了在协同训练中比较模型训练效果,基于这 2 个实体识别模型构建 2 个分类器,分别包括实体集、情感词典和情感规则 3 个部分。在协同训练的过程中,首先,从未进行实体标注和立场判定的数据集中抽取一定数量的数据,并利用这 2 个训练过的实体识别模型对未标注语料进行实体识别。同时,2 个实体识别模型通过判断微博文本中的实体类型,提取指定方向的实体,各形成 1 个实体集。然后,在这 2 个实体集的基础上,构造 2 个分类器,利用分类器分别结合情感词典和情感规则来判断微博立场,并对同一条博文进行倾向性分析。比较 2 个分类器得到的倾向性分析结果,判断样本置信度,从中选择高置信度(即 2 个分类器输出结果完全相同)的样本,并为此类样本添加指定类别的实体标记和倾向性标签,合并到已标记博文数据中。同时,对于低置信度的样本,采用主动学习模式,将分类器挑选出的分歧大(2 个分类器标注结果不同)的样本添加到已标注的数据集中。将更新扩充后的有标注数据集重新输入到协同训练模型,再次训练上述 2 个实体识别深度学习模型,不断迭代直到已标记的博文数据集达到足够的规模,进而获取最大的指定方向实体集。2.1 节将详细介绍微博实体识别模型及指定方向实体集的训练过程。

在基于情感词典和句法分析的微博倾向性分析部分,首先,模型基于学习得到指定方向实体集。然后,针对需要立场判断的微博数据,判断微博是否与指定方向的实体相关。如果微博文本中包含指定方向实体集内的任意实体,则对此类微博根据预先编写的情感规则和情感词典进行倾向性分析,实现对微博文本的立场判定。2.2 节将重点讨论 OASOSR 算法中基于情感词典和情感规则的倾向性分析。

2.1 基于协同训练和主动学习的实体识别模型

本节详细介绍基于协同训练和主动学习模型的构造细节、训练过程及微博情感对象实体集的提取流程。

协同训练假设数据拥有 2 个充分且条件独立的视图,即每个视图所包含的信息都可以支持生成最优学习器,且在给定类别标记的条件下 2 个视图相互独立。利用未标记数据在每个视图上都训练生成 1 个分类器,每个分类器选择置信度最高的样本生成伪标记,再将这些伪标记作为标记样本训练另 1 个分类器。本文对层叠隐马尔可夫模型(cascaded hidden Markov model, CHMM)和条件随机场(conditional random field, CRF)实体识别学习模型进行协同训练。

隐马尔可夫模型(hidden Markov model, HMM)是一种在自然语言处理领域中被广泛应用的统计模型。中文命名实体识别中的人名识别、地名识别、译名识别及机构名识别等都可以用 HMM 来解决。本文利用 CHMM 在统一的 HMM 中识别各类实体命名,自底向上分为人名识别 HMM、地名识别 HMM 和机构名识别 HMM 3 层。每一层 HMM 将产生的最好的若干个结果送到词图中供高层模型使用,最终实现人名、地名和机构名的实体识别。

CRF 模型是一种基于统计的判别式模型,结合了 ME 模型和 HMM 模型的特点。CRF 模型根据样本的特征生成预测,对所有的特征权重进行最优化,进而得到最优解,解决了 HMM 模型因独

立性假设而导致其不能考虑上下文特征的问题。在本文模型中,从经过分词的数据中挑选出包含在特定微博主题标签内的数据实体,形成候选词集。再利用 CRF 模型对候选词集中的候选实体进行实体抽取。

本文采用协同训练和主动学习的方法训练以上 2 个模型。主动学习方法具有在训练集较少的情况下能够获得较高分类准确率的优势,是对有监督学习方法的一种改进。有监督学习方法在学习过程中被动接受人工标注的样本集,通过将标注样本映射到目标函数空间来学习知识。而主动学习方法则试图从未标注的池中选择数量尽可能少的、机器无法正确标注的样本(低置信度的样本),并由领域专家进行标注后添加到下一次迭代过程中作为训练集,以提高分类的效率。主动学习通过选择质量高的少数样本进行学习以保证分类器的分类性能,同时,也可以减轻样本的标注复杂度。

情感对象实体集提取流程如图 2 所示,采用了半监督学习方法,通过协同训练和主动学习的方式对 2 个实体识别学习模型进行训练。在实体及情感类别已标注数据集中,有少量已标注微博文本分别标注了博文中包含的实体(实体 1、实体 2 等)和博文情感类别(正向、负向)。将有标注的微博文本进行切词处理,分别输入到 CHMM 和 CRF 实体识别学习模型,初步训练这 2 个模型。

在对 CHMM 和 CRF 实体识别学习模型进行协同训练的过程中,先从未标注实体及情感类别

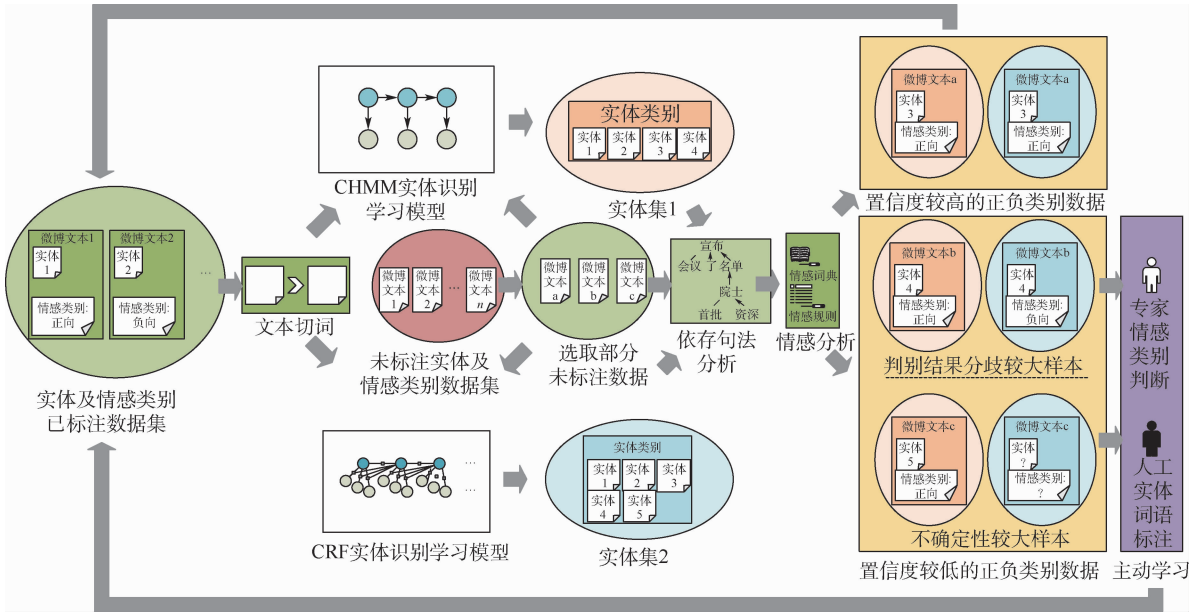


图 2 情感对象实体集提取流程

Fig. 2 Extraction flowchart of sentimental object entity sets

数据集中选取部分数据(微博文本 a、b、c),输入到经过初步训练的 CHMM 和 CRF 实体识别学习模型,提取微博中指定方向的实体,分别得到实体集 1 和实体集 2。分别基于 2 个实体集对相同的未标注数据进行依存句法分析,再基于给定的情感词典和情感规则进行情感分析,判断微博文本的情感类别。对于每条微博,都有基于实体集 1 和实体集 2 的 2 个标记结果,每个标记结果都包括微博所含有的实体(实体 n)和倾向(正向、负向、无法判断)。比较基于实体集 1 和实体集 2 得到的标记结果,针对同一微博文本,若 2 个分类器得到的结果中所标记实体和情感类别都完全相同,则判断该结果为置信度较高的正负类别数据。直接将这条微博文本和对应实体、情感类型标注添加到实体及情感类别已标注数据集;若 2 个结果中所标记实体或情感类别不完全相同,则判断该结果为置信度较低的正负类别数据,需要采用主动学习的方法处理。对于这类数据,针对同一微博文本,若 2 个分类器提取的指定方向实体完全相同,但情感类别相反,则认为该样本是分歧较大样本,需要将结果交由领域专家进行情感类别判断。若 2 个分类器提取的指定方向实体不同,或有 1 个分类器没有提取出指定方向实体,则认为该样本是不确定性较大样本,需要进行人工实体词语标注。同样地,将经由主动学习得到的微博文本和对应实体、情感类型标注添加到实体及情感类别已标注数据集,这样就完成了协同学习的第一轮循环。

对这一过程进行迭代,不断扩充实体及情感类别已标注数据集,直到已标记数据的数据量达到设定的停止阈值。获取该数据集中的所有标注实体所组成的实体集,作为 2.2 节中微博倾向性分析的基础。

2.2 基于情感词典和情感规则的微博倾向性分析

本节介绍基于情感词典和情感规则进行句法分析的微博立场判断算法。该算法通过情感规则识别微博中情感词所充当的句子成分,分析情感词与目标实体之间的修饰关系,从而判断微博的倾向性。然而,情感规则的制定依赖于依存句法分析。汉语语句以词语为基本单位,在词语之间拥有依存和支配关系。依存句法分析理论能够借用各级语言单位的依存关系,清晰地提炼出成分之间的修饰和搭配信息,从而达到解析语句的目的,进而总结句子成分之间的关联,制定情感规则。

基于情感词典和情感规则的微博倾向性分析

算法包含文档预处理、条件判断、实体信息判断、句法分析及规则判断 5 个步骤,输出结果为对应输入微博的博文倾向性类别。算法流程如图 3 所示。

首先,对原始博文进行数据清洗,经过简化繁体字、删除无效评论等操作后,得到有效数据,再使用结巴分词(数据来源:<https://github.com/fxsjy/jieba>),并去除分词结果中的停用词。然后,基

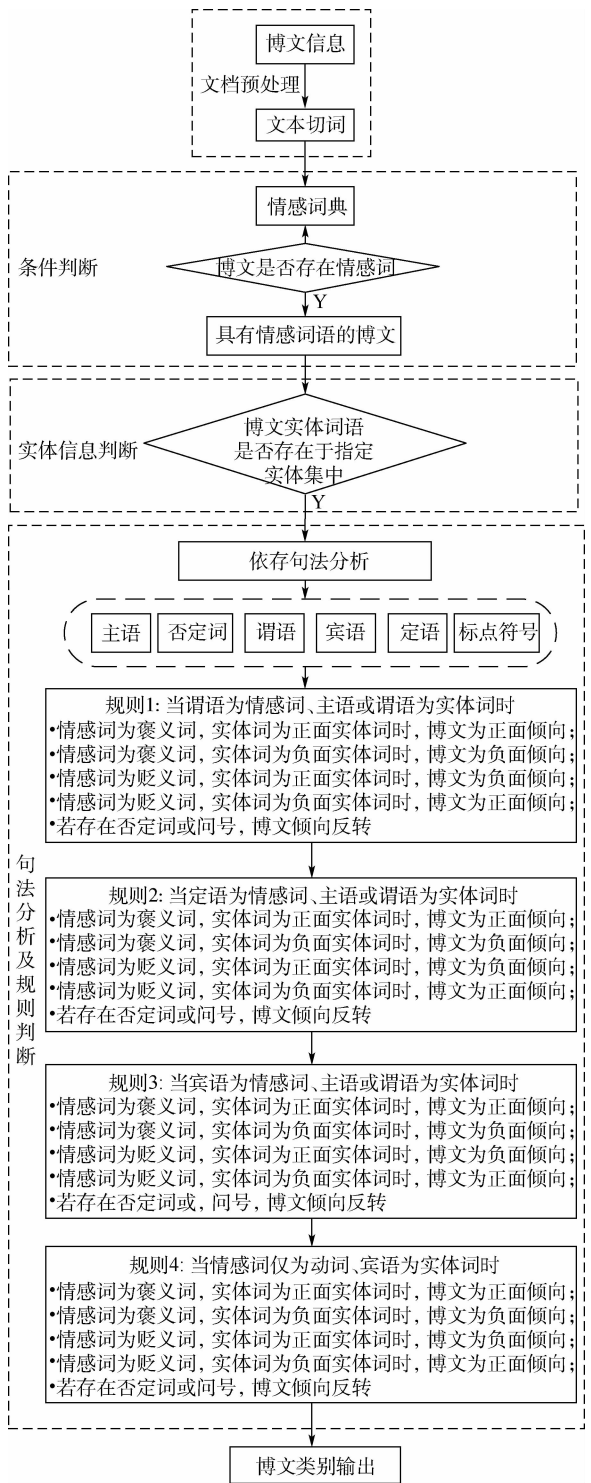


图 3 OASOSR 算法流程

Fig. 3 OASOSR algorithm flowchart

于情感词典,对经过处理的微博文本数据进行情感词判断。显然,只有包含情感词的微博文本才能表达倾向性,不包含情感信息的微博不具有立场分析的价值。因此,本文采用大连理工大学情感词本体库 DUTIR 情感词典(数据来源: <http://ir.dlut.edu.cn>)判断博文中是否存在情感词,筛选出含有情感词的微博文本,进行下一步处理。在含有情感词的微博文本基础上,进行实体信息判断,筛选出包含指定方向实体的微博。实体信息判断基于 2.1 节中通过协同训练和主动学习得到的实体集完成。最后,提取微博文本中实体词语,若实体存在于该指定方向的实体集之中,则筛选出这条微博,对其中的句子进行依存句法分析。通过识别句子的主语、否定词、谓语、宾语、定语、标点符号等成分,基于句法分析、情感规则将句子分为 4 种不同的类型:①当谓语为情感词、主语或谓语为实体词时;②当定语为情感词、主语或谓语为实体词时;③当宾语为情感词、主语或谓语为实体词时;④当情感词仅为动词、宾语为实体词时。

情感词在句子中充当不同成分时,具体分类规则如 OASOSR 算法所示(见图 3)。经过对应的情感规则判断,最终输出该微博文本的立场。

- 算法 1 OASOSR 算法。
- 输入:微博文本。
- 输出:博文类别。
- 步骤 1 文档预处理。文本切词。
- 步骤 2 条件判断。依据情感词典判断博文是否存在情感词,若存在情感词,则对句子进行下一步处理。
- 步骤 3 实体信息判断。提取博文中实体词语,若实体存在于指定实体集之中,则进行下一步处理。
- 步骤 4 句法分析及规则判断。
- 步骤 4.1 对句子进行依存句法分析,识别句子的主语、否定词、谓语、宾语、定语、标点符号等成分。

- 步骤 4.2 设定情感判断规则。
- 1) 规则 1:当谓语为情感词、主语或谓语为实体词时:
- ① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。
- ② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。
- ③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。
- ④ 情感词为贬义词,实体词为负面实体词

- 时,博文为正面倾向。
- ⑤ 若存在否定词或问号,博文倾向反转。
- 2) 规则 2:当定语为情感词、主语或谓语为实体词时:
- ① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。
- ② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。
- ③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。
- ④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。
- ⑤ 若存在否定词或问号,博文倾向反转。
- 3) 规则 3:当宾语为情感词、主语或谓语为实体词时:
- ① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。
- ② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。
- ③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。
- ④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。
- ⑤ 若存在否定词或问号,博文倾向反转。
- 4) 规则 4:当情感词仅为动词、宾语为实体词时:
- ① 情感词为褒义词,实体词为正面实体词时,博文为正面倾向。
- ② 情感词为褒义词,实体词为负面实体词时,博文为负面倾向。
- ③ 情感词为贬义词,实体词为正面实体词时,博文为负面倾向。
- ④ 情感词为贬义词,实体词为负面实体词时,博文为正面倾向。
- ⑤ 若存在否定词或问号,博文倾向反转。
- 步骤 4.3 输出博文的立场。

在 OASOSR 算法中,文本切词的时间复杂度为 $O(n^2)$,条件判断、实体信息判断的时间复杂度为 $O(n)$,而句法分析的时间复杂度为 $O(Kn^3)$,可以得知 OASOSR 算法时间复杂度为 $O(Kn^3)$ 。

3 实 验

3.1 实验设置

为证明 OASOSR 算法的实用性,本文实现了 OASOSR 算法的自对比实验和他比实验。本文通过网络爬虫获取了 38 175 条新冠肺炎疫情舆论

微博数据(数据来源: <https://github.com/Zu-Qin131/Weibo-public-opinion-analysis>),并对新冠肺炎疫情舆情言论微博数据进行人工实体标注,分为正面和负面情感博文,用于自对比实验;并选用 weibo_senti_100k(数据来源: <https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/datasets>)的开源微博情感数据集,从中选取部分数据进行实体标注和人工正面、负面情感标注,应用于自对比实验和他比实验。上述数据集的特点总结如表 3 所示。

表 3 对比实验中选取的数据集

Table 3 Datasets selected for comparative experiment			
数据类型	数据集	大小/条	数据格式
爬取数据	#新冠肺炎疫情#话题	38 175	标签(正向、负向),文本
公开数据	weibo_senti_100k 开源微博情感	119 989	标签(正向、负向),文本

实验中构造的微博情感词典是在大连理工大学情感词汇本体库 DUTIR 的基础上,添加与新冠肺炎疫情舆论相关的情感词得到的。最终构建的情感词典共有 23 651 个中文情感词,包括 9 833 个正面情感词和 13 818 个负面情感词。

为了测试在数据量变化情况下 OASOSR 算法的稳定性、鲁棒性与准确率变化趋势,本节设计了基于 OASOSR 算法的自对比实验。如表 4 所示,从新冠肺炎疫情话题数据集中随机选取 485 条新冠肺炎疫情舆情言论微博数据,进行人工实体标注,并将博文分为正面情感博文和负面情感博文,其中已标注正面情感博文 331 条,负面情感博文 154 条,作为小规模数据集。

同时,也在 weibo_senti_100k 的开源微博情感数据集上使用该方法进行实验,这是一个开源的大规模数据集,共有 119 989 条微博情感数据,从中随机选取 4 000 条微博数据进行数据标注,并作为大规模数据集,其中标注了正面情感博文 2 200 条,负面情感博文 1 800 条。

为证明 OASOSR 算法在不同规模数据集上的有效性,设计了 4 组实验,在不同规模下对 OASOSR 算法的鲁棒性进行了测试。首先,在小规模

表 4 对比实验中选取的数据集大小

Table 4 Size of datasets selected for comparative experiment	
数据集	实验数据条数
#新冠肺炎疫情#话题	485
weibo_senti_100k 开源微博情感	4 000

数据集(新冠肺炎疫情话题数据集)上进行实验,从新冠肺炎疫情话题数据集中随机选取 200 条作为训练集,剩余的 285 条博文作为测试集。然后,在大规模数据集上进行实验,从 weibo_senti_100k 开源微博情感数据集中选取了 2 000 条作为训练集,剩余 2 000 条作为测试集进行实验。在 2 个数据集上分别进行 4 组实验,每组实验分别从训练集中选取 25%、50%、75% 和 100% 的数据用于训练,生成训练结果,并对比观察 2 个模型在测试集上情感分类的正确率,正确率对比如图 4 所示。

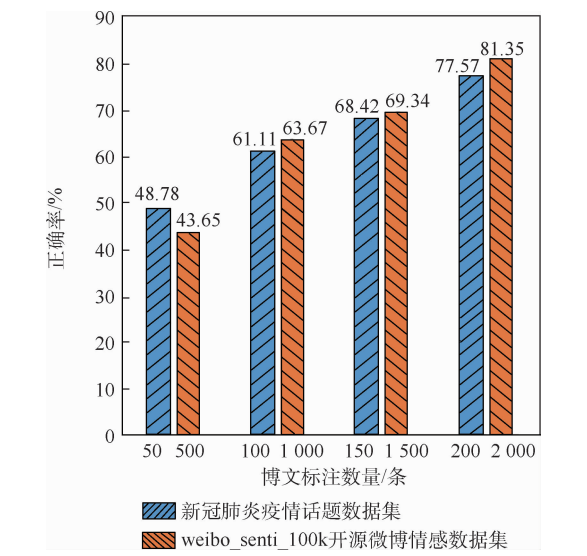


图 4 不同规模数据集上 OASOSR 算法立场判断正确率
Fig. 4 Accuracy of standpoint judgement by OASOSR algorithm on different datasets

在小规模数据集上,当实体标注博文的数量为 50 条时,标注数据过少,识别出的情感对象实体较少,因此博文情感识别率较低。当实体标注博文的数量为 100、150、200 条时,正面和负面情感博文的判断正确率持续提高。相似地,在大规模数据集上,当实体标注博文的数量为 500 条时,正面和负面情感博文的判断正确率都比较低,而当情感标注博文的数量为 1 000、1 500、2 000 条时,正面和负面情感博文的判断正确率持续提高。

可以看出,随着训练数据中标注实体的博文数量增加,在不同规模的数据集上,算法对微博立场的判断正确率都在逐渐提高。OASOSR 算法在数据集规模变化差别较大的情况下,仍具有很好的鲁棒性。

3.2 他比实验

他比实验选取 weibo_senti_100k 开源微博情感数据集作为实验语料。为了验证 OASOSR 算法的有效性,本文选取了基于 SVM 的有监督情感文本分类方法(sentiment classification of texts

based on SVM, SCSVM)^[22] 和基于极性词典情感分析的无监督情感分类方法 (sentiment analysis method based on a polarity lexicon, SAMPL)^[23], 分别与本文算法进行对比。在 SCSVM 方法中, 先构建情感词典, 并进行情感特征选取及情感特征加权, 再使用 SVM 分类的方法对文本进行情感识别及分类。SAMPL 方法则综合了基础词典、领域词典、网络词典及修饰词词典, 将极性词和修饰词组合成极性短语, 以极性短语为极性计算的基本单元, 进行情感分析。

表 5 展示了实验中基于上述 2 个模型的微博立场判断正确率和基于 OASOSR 算法的微博立场判断正确率。对比表 5 中正确率可知, 在同等数量的训练集和测试集中, OASOSR 算法要优于 SCSVM 方法和 SAMPL 方法。这是因为: 本文提出的 OASOSR 算法充分考虑了对情感对象的识别, 微博情感识别更有针对性, 在立场判断任务上的正确率相对较高。由对比实验结果可以证明, 相较于有监督学习方法和无监督学习方法, 半监督学习方法更适用于指定实体的博文立场判断。

表 5 基于不同模型的微博立场判断正确率

Table 5 Accuracy of Weibo standpoint judgement based on different models

方法	正确率/%
SCSVM	78.56
SAMPL	71.35
OASOSR	81.35

3.3 规则分析实验

为证明 OASOSR 算法中不同情感规则在微博文本文倾向性分析中会起到作用, 本文设计了 2 组规则分析实验。在第 1 组实验中, 设置变量为仅保留情感规则 1~4 条中的 1 条, 去除其他 3 条规则; 在第 2 组实验中, 设置变量为去除情感判断规则 1~4 条中的 1 条, 按原有顺序保留其余 3 条规则。在新冠肺炎疫情话题数据集和 weibo_senti_100k 开源微博情感数据集上进行这 2 组实验, 生成实验结果, 并将结果与去除判断规则前的实验结果进行对比。图 5 中, “单一” 代表仅保留某 1 条规则, 去除其他 3 条情感规则得到的实验正确率; “缺失” 表示缺失某 1 条规则, 但保留其他 3 条规则的得到的实验正确率。从实验结果对比可以看出, 在去除某 1 条情感规则后, 立场判定的正确率均低于原有算法模型。特别是去除规则 1 后, 正确率下降了 34.36%。因此, OASOSR 算法中 4 条规则互相支撑且都具有必要性, 缺一不可。

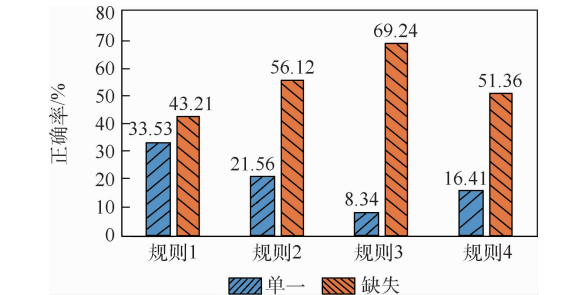


图 5 基于不同模型筛选条件的微博立场判断正确率
Fig. 5 Accuracy of Weibo standpoint judgement based on filtering conditions of different models

4 结 论

微博情感的分类较为简单, 无法判断微博对特定方向实体的立场问题。针对微博文本数据信息量大、情感倾向性强的特点, 本文提出了使用 OASOSR 算法解决微博立场判定问题。该算法结合深度学习与情感规则, 实现微博立场判定。首先, 本文采用半监督学习的方法, 分别基于 CHMM 和 CRF 模型, 对微博文本进行协同训练, 再通过主动学习的方法, 提取微博文本中的指向性实体集, 并构造了 2 个分类器。然后, 通过分析大量微博文本中的句法特征, 构建情感规则, 提取句子的主成分, 去除与立场判断不相关的信息, 将表达方式较为随意的微博文本规范化。基于经过处理的博文数据, 综合判断指向性实体的正面和负面性, 结合博文句子中情感词的褒贬义、分析情感词充当的句子成分及情感词修饰的实体类型, 进而分析博文的倾向性, 确定微博对指定主题所表达的立场。为证明 OASOSR 算法的实用性, 本文实现了自对比实验和他比实验。自对比实验显示, 随着标注实体的博文数量增加, 模型对博文立场判断的正确率持续提升, 证明了该算法可以有效增强立场判断效果。将 OASOSR 算法与 SCSVM 方法、SAMPL 方法进行对比实验, 在不同标注数量博文上进行对比的实验结果显示, OASOSR 算法判断博文立场的正确率显著高于对比方法。

参考文献 (References)

[1] GIACHANOU A, MELE I, CRESTANI F. Explaining sentiment spikes in twitter[C] // Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2016: 2263-2268.

[2] 王志涛, 於志文, 郭斌, 等. 基于词典和规则集的中文微博情感分析[J]. 计算机工程与应用, 2015, 51(8): 218-225.

WANG Z T, YU Z W, GUO B, et al. Sentiment analysis of Chinese micro blog based on lexicon and rule set[J]. Computer Engineering and Applications, 2015, 51(8): 218-225 (in Chi-

nese).

[3] 王灿伟. 基于主题提取的海量微博情感分析[J]. 南京大学学报(自然科学),2017,53(3):549-556.

WANG C W. Sentimental analysis of massive micro-blog based on topic extraction[J]. Journal of Nanjing University (Natural Sciences),2017,53(3):549-556(in Chinese).

[4] EBRAHIMI J, DOU D J, LOWD D. A joint sentiment-target-stance model for stance classification in tweets[C]//Proceedings of the 26th International Conference on Computational Linguistics,2016:2656-2665.

[5] PAK A, PAROUBEK P. Twitter as a corpus for sentiment analysis and opinion mining[C]//Proceedings of International Conference on Language Resource and Evaluation,2010:13-20.

[6] PANG B, LEE L, VAITHYANATHAN S, et al. Thumbs up?: Sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 Conference on Empirical Methods on Natural Language Processing. New York: ACM,2002:79-86.

[7] 莫雨洁,金琴,吴慧敏. 基于多文本特征融合的中文微博的立场检测[J]. 计算机工程与应用,2017,53(21):77-84.

DIAN Y J, JIN Q, WU H M. Stance detection in Chinese microblogs via fusing multiple text features[J]. Computer Engineering and Applications,2017,53(21):77-84(in Chinese).

[8] 李俭兵,刘栗材. 基于改进型神经网络的影评文本情感分析算法[J]. 计算机工程与科学,2019,41(12):2261-2269.

LI J B, LIU S C. A film criticism sentiment analysis algorithm based on improved neural network[J]. Computer Engineering and Science,2019,41(12):2261-2269(in Chinese).

[9] LI D, QIAN J. Text sentiment analysis based on long and short term memory[C]//2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI). Piscataway: IEEE Press,2016:471-475.

[10] 张仰森,郑佳,黄改娟,等. 基于双重注意力模型的微博情感分析方法[J]. 清华大学学报(自然科学版),2018,58(2):122-130.

ZHANG Y S, ZHENG J, HUANG G J, et al. Microblog sentiment analysis method based on a double attention model[J]. Journal of Tsinghua University (Science and Technology),2018,58(2):122-130(in Chinese).

[11] 朱晓光,聂培尧,林培光. 基于监督学习的微博情感分类方法[J]. 计算机应用与软件,2015,32(8):238-242.

ZHU X G, NIE P Y, LIN P G. Supervised learning based on microblogging sentiment classification method[J]. Computer Applications and Software,2015,32(8):238-242(in Chinese).

[12] 段吉东,刘双荣,马坤,等. 基于集成学习的文本情感分类方法[J]. 济南大学学报(自然科学版),2019,33(6):483-488.

DUAN J D, LIU S R, MA K, et al. Text sentiment classification method based on ensemble learning[J]. Journal of University of Jinan(Science and Technology),2019,33(6):483-488(in Chinese).

[13] TURNEY P D. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,2002:417-424.

[14] BLOOM K, ARGAMON S. Automated learning of appraisal extraction patterns[J]. Language and Computers,2010,71(2):249-260.

[15] GUO J L, PENG J E, WANG H C. An opinion feature extraction approach based on a multidimensional sentence analysis model[J]. Cybernetics and Systems,2013,44(5):379-401.

[16] AGRAWAL A, XIE B, VOVSHA I, et al. Sentiment analysis of Twitter data[J]. International Journal of Computer Applications,2013,139(11):880-887

[17] CAMBRIA E, PORIA S, HAZARIKA D, et al. Sentinet5: Discovering conceptual primitives for sentiment analysis by means of context embeddings[C]//32nd AAAI Conference on Artificial Intelligence,2018:1795-1802.

[18] DANDAPAT S. Handbook of natural language processing(second edition)[J]. Machine Translation,2011,25(4):377-381.

[19] SINDHWANI V, MELVILLE P. Document-word co-regularization for semi-supervised sentiment analysis[C]//18th IEEE International Conference on Data Mining. Piscataway: IEEE Press,2008:1025-1030.

[20] LIU Z, DONG X, GUAN Y, et al. Reserved self-training: A semi-supervised sentiment classification method for Chinese micro-blogs[C]//Proceedings of LJCNP,2013:455-462.

[21] SCUDDER H. Probability of error of some adaptive pattern-recognition machines[J]. IEEE Transactions on Information Theory,1965,11(3):363-371.

[22] 陈培文,傅秀芬. 采用 SVM 方法的文本情感极性分类研究[J]. 广东工业大学学报,2014,31(3):95-101.

CHEN P W, FU X F. Research on sentiment classification of texts based on SVM[J]. Journal of Guangdong University of Technology,2014,31(3):95-101(in Chinese).

[23] 张成功,刘培玉,朱振方,等. 一种基于极性词典的情感分析方法[J]. 山东大学学报,2012,47(3):47-50.

ZHANG C G, LIU P Y, ZHU Z F, et al. A sentiment analysis method based on a polarity lexicon[J]. Journal of Shandong University,2012,47(3):47-50(in Chinese).

Weibo tendency analysis based on sentimental object
recognition and sentimental rules

WANG Zechen¹, WANG Shupeng^{1,*}, SUN Liyuan², ZHANG Lei¹, WANG Yong¹, HAO Bingchuan¹

(1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100193, China;

2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100085, China)

Abstract: Weibo contains a large number of information reflecting users’ likes and dislikes, which is important for popular trend judgment, precision marketing, public opinion monitoring, etc. However, the existing methods tend to focus on the classification of Weibo sentiment. In order to solve the problem of Weibo tendentiousness analysis and position detection, we employ semisupervised learning method, through collaborative training and active learning. We train entity recognition models and combine deep learning with emotional rules. Moreover, the sentiment rules based on principal component analysis are constructed to extract the main components of sentences, normalize the spoken text into the specified format. Then we use the positive and negative aspects of directional entities, the positive and negative meanings of emotional words, and the sentence components of emotional words to judge the tendency of blog posts, and conduct deeper analysis on position classification. Finally, the self comparison experiment and other comparison experiment on different scale data sets show that with the increase of the number of blog posts of labeled entities, the accuracy of the model continues to improve, and the accuracy of this method is significantly higher than the comparison method, which is 2.79% and 10.00% higher than the existing research methods.

Keywords: sentiment analysis; position detection; semi-supervised learning; tendency; sentimental rule; co-training; active learning

Received: 2020-08-09; **Accepted:** 2020-09-25; **Published online:** 2021-03-09 20:12
URL: kns.cnki.net/kcms/detail/11.2625.V.20210309.0928.002.html
Foundation item: National Natural Science Foundation of China (61931019)
*** Corresponding author.** E-mail: wangshupeng@iie.ac.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0557

基于能源分解的用户用电行为模式分析

卢瑞瑞, 于海阳, 杨震*, 赖英旭, 杨石松, 周明

(北京工业大学 信息学部, 北京 100124)

摘 要: 随着智能电网的普及和大数据技术的发展,利用用电数据分析用户的用电行为越来越受到关注,现存的能源分解方法无法满足实际应用中对分辨率和分解准确率的高要求,以及聚类分析方法过于粗糙没有充分挖掘每类电器的用电特点。提出了基于能源分解的用户用电行为分析方法。在判别式稀疏编码算法模型的基础上,针对 L_0 正则项不易求解、 L_1 正则项稀疏约束效果不理想的问题,提出用 $L_{1/2}$ 正则项稀疏约束进行能源分解,并且把用户之间的同质性作为正则项加入基础模型来修正模型的性能。基于能源分解的结果,使用用户单类电器的用电特征代替总用电特征精细化分析用户的用电行为,并改进传统的 K-Mean 聚类算法进行实验验证。实验结果表明:所提出的基于 $L_{1/2}$ 正则项稀疏约束和同质性约束的能源分解方法相比于传统判别式稀疏编码算法,能够有效提升能源分解的准确率。同时,基于能源分解的用户用电行为聚类分析效果也有明显提升。

关 键 词: 智能电网; 能源分解; 聚类分析; 非负矩阵分解; 同质性关系

中图分类号: V221+.3; TB553

文献标志码: A **文章编号:** 1001-5965(2022)02-0311-13

近年来,大数据技术给人们的生活带来了许多便利,而智能电网被看作是大数据技术的应用之一,吸引了很多学者利用用户数据挖掘数据中潜在价值方面的研究。

能源问题是当今社会面临的最大挑战之一。在美国,建筑消耗的能源占总能耗的 40%,其中 73% 是电力消耗^[1]。研究表明,有效的能源管理可以减少建筑物的 10% ~ 15% 电力消耗^[2],用户对家用电器用电量的掌握有助于调节用电行为,实现能源优化调度。而能源分解可以获得用户的详细用电信息,让用户对自己的用电状态更加了解,在一定程度上也可以诱导用户用电行为达到节能的目的。另外,能源分解后的数据可以反映详细的能源使用情况,给用户提供更改进的依据。详细的能源分解可以实现对家庭用户电力消费行

为分析,实现配电的双向交互、信息交换、信息处理等环节的智能控制,提高供电侧的可靠性和安全性,从而实现用户方和供电方的利益最大化。

由于智能电网是直接面向用户的,网络结构复杂、类型繁多,如何深入分析用户的用电行为,挖掘出用户用电行为潜在价值,成为该领域的研究重点和难点。20 世纪 80 年代,麻省理工学院提出了非侵入式负载监测(NILM)^[3]方法,该方法首次从智能电表端采集用户的聚合用电信号并从中推断出每个电器的用电信号。随后,研究人员在这些方法的基础上寻找更有效的电信号特征,如稳态功率的谐波、电流消耗和瞬态噪声^[4-7],这些方法都称为基于事件的方法,其依赖于通过捕获特征来识别设备^[8];另外一种方法不依赖于事件检测分类设备,而是综合考虑所有的

收稿日期: 2020-09-27; 录用日期: 2020-11-06; 网络出版时间: 2021-03-10 10:33
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210309.1030.004.html
基金项目: 国家自然科学基金(61671030); 北京市长城学者计划(CIT&TCD20190308); 中国博士后科学基金(2019M660377)
* 通信作者: E-mail: yangzhen@bjut.edu.cn

引用格式: 卢瑞瑞, 于海阳, 杨震, 等. 基于能源分解的用户用电行为模式分析[J]. 北京航空航天大学学报, 2022, 48(2): 311-323. LU R R, YU H Y, YANG Z, et al. User electricity consumption behavior mode analysis based on energy decomposition[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 311-323 (in Chinese).

样本建立预测模型,是基于非事件的方法。2010 年,Kolter 等^[9]使用稀疏编码算法来学习每个设备在一周内的功耗模型,再结合这些学习模型,仅使用其聚合信号来预测以前未知家庭中不同设备的功耗。Kim^[10]和 Parson^[11]等使用因子马尔可夫模型(FHMM)将用户的聚合用电信号分解成单个电器的用电信号,并且隐藏设备的状态。2012 年,Kolter 和 Jaakkola^[12]在 FHMM 框架的基础上,提出了基于凸规划的近似算法。随着深度学习的发展,一些研究人员提出使用深度神经网络的方法通过分类来识别设备从而进行能源分解^[13-15]。除了这些单通道源分离工作外,研究人员还致力于利用其他信息来进行能量分解。2016 年,Batra 等^[16]提出了单个设备的功耗可以通过具有相似特征家庭的数据估计得出。然而,所有这些技术大部分都需要高分辨率的数据,而大多数智能仪表的采样率都达不到这个要求。另外,这些方法需要复杂且耗时的分解过程,难以在普通家庭中快速投入使用。

2005 年,李培强等^[17]基于模糊聚类原理提出了模糊 C 均值算法和模糊等价关系 2 种算法对变电站的综合负荷进行分类的方法。2009 年,王璨和冯勤超^[18]提出从当前的市场价值、潜在市场价值和区域贡献价值等角度来对用户的用电行为进行聚类研究。这 2 种聚类角度都过于宏观,未考虑到用户的用电负荷特征,不能实现用户电力负荷的精细化管理。2010 年,李欣然等^[19]提出使用模糊 C 均值算法对用户的日用电负荷曲线进行分析来对用户所属的行业进行分类。2018 年,Zhong 等^[20]提出使用 K-Means 算法对用户添加标签,实现用户画像为电力公司了解用户的电力消耗习惯、了解用户需求、提高服务质量提供数据支撑。2019 年,Nordahl 等^[21]提出使用 K-Me-doide 方法来分析和理解家庭的用电量数据,通过提取的知识为每个特定家庭创建正常用电行为模型用于异常行为检测。

综上分析,在众多的能源分解方法中,稀疏编码方法相对简单有效,根据相似用户预测能耗的方法有一定的合理性但是不完全据此预测,而智能电网用户用电行为分析方法仍过于粗糙,需要进一步精细化分析。如何提高能源分解的准确性,并合理利用分解结果对用户的用电行为进行分析是一个有趣且充满挑战的问题。

本文针对现有能源分解系统分解准确率不高、难以应用到普通用户、用户用电行为聚类分析特征欠缺等问题,首先,提出基于稀疏约束的能源

分解方法,在不丢失信号特征的前提下更简单地获取信号中蕴含的主要信息,同时也方便对信号做进一步的加工处理。其次,提出基于同质性约束的能源分解方法,在能源分解的过程中把用户之间同质性考虑进去,来提高分解系统的准确性。最后,通过对基于用电模式的用户行为聚类分析,完成智能电网用户的分类,并分析每类用户的用电特点。

1 系统描述

1.1 系统架构流程

本文的整体结构可以表述为如图 1 所示的四元结构图,普通用户家庭、智能电网、数据分解系统和聚类分析系统是整个结构的 4 个组成部分。普通用户家庭从智能电网获取电能,从数据分解系统获取一段时间内的用电反馈。智能电网为用户提供电能,并采集用户的用电数据反馈给数据分解系统和聚类分析系统。数据分解系统将总用电数据分解成单类用电器的用电数据反馈给用户,或者提供给聚类分析系统分析用户的用电行为。聚类分析系统通过数据分解系统提供用户的详细用电数据对用户用电行为进行聚类分析应用于网络规划、需求响应和信息推荐等。

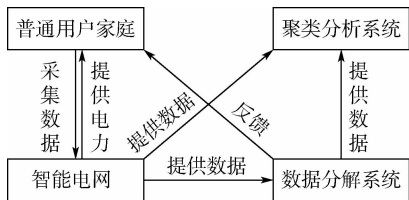


图 1 系统四元结构

Fig. 1 System quaternion structure

系统首先从智能电网采集数据集,对数据集进行筛选和预处理,选择用电数据多的用户,统一数据格式,把数据处理成结构化的矩阵数据为使用做准备;其次能源分解模块把用户单类电器用电数据作为训练集训练模型,用户总的用电数据作为测试集,把用户总的用电数据分解成单类用电器的用电数据输出;然后评测能源分解模块的性能,能源分解模块分解出来的单类用电器的用电数据也可以提供给其他第三方应用;最后使用能源分解前的总用电数据和能源分解后的单类电器用电数据对用户进行聚类,并对比分析其聚类结果。

1.2 系统主要模块

1.2.1 数据预处理模块

本文采用 Pecanstreet 数据集进行实验,该数据集中包含 69 个电器,每个用户的单类电器是每小时采集一次用电量,如果采集 500 个家庭一

个月的用电量可达到 2 亿多条数据。数据规模庞大,用户用电器类别不一致,并且包含很多无效的数据。因此,本文对原始数据集进行预处理。首先,筛选可用的数据,选择部分用电器和用户的用电数据进行研究;其次,处理缺失数据,对邻近时间段的数据求均值进行插入来完善数据集;再次,处理异常数据,把长时超出电器负荷的值作为异常数据,用邻近时间段的数据来替代;最后,把数据转换成本文需要的数据格式为实验做准备。其中,数据筛选部分选取数据的准则为:对于电器类别,选择大部分家庭中都有的用电器作为主要研究对象,不常见或者只存在部分家庭中的用电器的用电数据对其求和作为杂类;对于用户,选择用电数据较多的用户,丢弃连续多天不用电或用电异常的用户。最终得到 5 类电器(空调、电炉、洗衣机、冰箱和其他电器)的总和,264 个用户 2 周内的用电数据用于本文的实验。

1.2.2 能源分解模块

本文的能源分解模块把用户的总用电数据分解成单类用电器的用电数据。非负矩阵分解(nonnegative matrix factorization, NMF)是处理盲源分离问题的一种有效方法,其可以通过一些线性组合来近似混合信号。2010 年, Kolter 等^[9]提出判别式稀疏编码的能源分解方法,认为用电数据存在稀疏性,在用非负矩阵进行能源分解时,优化目标中加入了 L_1 正则化项来进行稀疏约束。本文基于稀疏非负矩阵分解这个研究方向来改进原有的算法,提升能源分解的准确率。首先,收集数据并对数据预处理,把数据稀疏表示,生成稀疏约束正则项 L_1 ;其次,计算不同家庭特征和用电量之间的相关性进行相关性分析,建立同质性正则项 L_2 ;然后,基于非负矩阵分解算法、同质性约束和稀疏约束建立模型;最后,继续优化模型并进行模型求解。

1.2.3 聚类分析模块

本文的聚类分析模块利用能源分解模块得到的分解结果对用户的用电行为进行聚类分析。首先,总结了传统的智能电网用户聚类方法,然后,选择基于 K-Means 聚类算法的智能电网用户聚类方法对其改进应用于本文提出的方法。

2 能源分解和聚类分析设计

2.1 基于 $L_{1/2}$ 范数和同质性约束的能源分解方法

2.1.1 基于非负矩阵分解的能源分解模型

NMF 是由 Lee 和 Seung^[22] 于 1999 年在 *Nature*

上提出的一种矩阵分解方法^[23],该方法分解后的所有分量都是非负的,并且可以实现降维的目的。具体地,NMF 方法主要解决如下问题:给定非负矩阵 X ,找到非负矩阵 B 和 A 使得 $X \approx BA$,目标函数可以建模为

$$\min_{B,A} \|X - BA\|_F^2 \quad \text{s. t. } B \geq 0, A \geq 0 \tag{1}$$

式中: X 为待分解矩阵; B 为特征矩阵,每一列是一个基向量即特征,矩阵 A 是代表特征重要程度的权重矩阵,即 A 的行表示每个基向量的贡献大小。优化式(1),找到最优的 B 和 A 使得式(1)最小。式(1)在求解时,如果只考虑 B 或者 A 一个变量时是凸函数,但是同时考虑 2 个变量时就是非凸问题,对于非凸函数想要找到全局最优解是不切实际的。Lee 和 Seung^[22] 提出使用乘性迭代法进行求解,对 2 个变量交替迭代寻找最优,迭代如下:

$$B_{ij} \leftarrow B_{ij} \frac{(XA^T)_{ij}}{(BAA^T)_{ij}} \tag{2}$$

$$A_{ij} \leftarrow A_{ij} \frac{(B^T X)_{ij}}{(B^T B A)_{ij}} \tag{3}$$

对于 NMF 方法进行优化求解时,是一个非凸优化问题,最终得到的是一个局部最优解,也就是说基本 NMF 方法的最优解不唯一。因此,在实际应用中,研究者通常会加入一些数据的先验知识,缩小解的范围并且使求解结果更加合理。而这些先验知识则是通过给 NMF 添加正则项约束的方法加入,添加正则化后的 NMF 方法模型如式(4)所示:

$$\min_{B,A} \|X - BA\|_F^2 + \alpha \Omega_1(B) + \beta \Omega_2(A) \tag{4}$$

式中:函数 Ω_1 和函数 Ω_2 表示正则化项,对矩阵 B 和矩阵 A 进行约束; α 和 β 为正则项系数,来控制该正则项约束的强弱。正则化的 NMF 方法因其加入了先验知识,使矩阵分解的结果具有了一定的实际意义,更具有可解释性,因此越来越受到关注。本文提出的能源分解方法即建立在正则化的 NMF 方法基础之上。

盲源分离是从观测到的混合信号中分离出原始信号,混合信号通常来自于多个传感器的输出,并且这多个传感器的输出信号都是独立的。能源分解问题在一定程度上可以理解为盲源分离,给定总的用电量信号,分离出单类用电器的用电信号。NMF 是处理盲源分离问题的一种有效方法,其通过线性组合来近似混合信号^[24]。2010 年, Kolter 等^[9] 从盲源分离技术中受到启发,提出使

用非负稀疏编码的方法来解决能源分解问题。假设有 k 类不同的用电器,如电视、冰箱、加热器等,对每类电器 $i = k$ 有矩阵 $\mathbf{X}_i \in \mathbf{R}^{m \times l} (i = 1, 2, \cdots, k)$,其中 $\mathbf{X}_i$ 的每一列表示特定用户特定设备 1 周内的用电数据,数据时间间隔为 1 h。例如:定义 $\mathbf{X}_1^j$ 为矩阵 $\mathbf{X}_1$ 的第 j 列,可能表示 1 个用户 1 周内冰箱的能耗, $\mathbf{X}_2^j$ 则表示相同用户同一时间段内洗衣机的能耗。同时,定义用户所有电器总的能耗为 $\bar{\mathbf{X}} = \sum_{i=1}^k \mathbf{X}_i$, $\bar{\mathbf{X}}$ 的第 j 列包含给定用户同一时间段内所有用电器的能源消耗。在训练过程中,假设可以通过智能电表获取每类用电器的能耗矩阵 $\mathbf{X}_1, \cdots, \mathbf{X}_k$,然而在测试过程中,假设只能通过智能电表获取总的用电数据矩阵 $\bar{\mathbf{X}}'$,算法的最终目标是分离总的用电数据矩阵得到每类电器的用电数据矩阵 $\mathbf{x}'_1, \mathbf{x}'_2, \cdots, \mathbf{x}'_k$ 。非负稀疏编码的能源分解方法,首先为每类单独的用电器训练单独的模型,然后使用训练出来的模型来分离总的聚合信号。稀疏编码方法使用近似值 $\mathbf{X}_i \approx \mathbf{B}_i \mathbf{A}_i$ 来为每类单独的电器建模,其中 $\mathbf{B}_i \in \mathbf{R}^{m \times n}$ 的每一列都表示一个基,也叫做特征, $\mathbf{A}_i \in \mathbf{R}^{m \times n}$ 的列表示每个基的激活系数。稀疏编码方法给系数矩阵加稀疏约束,使矩阵 $\mathbf{A}_i$ 中有更多的零值。具体的,可以为每类电器通过式(5)建模:

$$\begin{aligned} \min_{\mathbf{A}_i \geq 0, \mathbf{B}_i \geq 0} \quad & \frac{1}{2} \|\mathbf{X}_i - \mathbf{B}_i \mathbf{A}_i\|_F^2 + \lambda \sum_{p,q} (\mathbf{A}_i)_{pq} \\ \text{s. t.} \quad & \|b_i^{(j)}\|_2 \leq 1 \quad j = 1, 2, \cdots, n \end{aligned} \tag{5}$$

式中: $\sum_{p,q} (\mathbf{A}_i)_{pq}$ 为稀疏约束正则项; $b_i^{(j)}$ 为矩阵 $\mathbf{B}_i$ 中的第 j 列; λ 为正则项系数用来控制稀疏约束的强弱。使用式(5)为每类单独的用电器训练出 $\mathbf{B}_i$ 和 $\mathbf{A}_i$,再使用训练结果去分解一个新的聚合信号 $\bar{\mathbf{X}}$ 。具体的,已知每类电器的信号特征,通过式(6)为每类电器找到最优的系数矩阵 $\hat{\mathbf{A}}_i$ 来重构聚合信号。

$$\begin{aligned} \hat{\mathbf{A}}_{1,k} = \arg \min_{\mathbf{A}_{1,k} \geq 0} \quad & \left\| \bar{\mathbf{X}} - [\mathbf{B}_1 \cdots \mathbf{B}_k] \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_k \end{bmatrix} \right\|_F^2 + \\ & \lambda \sum_{i,p,q} (\mathbf{A}_i)_{pq} = \arg \min_{\mathbf{A}_{1,k} \geq 0} F(\bar{\mathbf{X}}, \mathbf{A}_{1,k}, \mathbf{B}_{1:k}) \end{aligned} \tag{6}$$

式中: $[\mathbf{B}_1 \cdots \mathbf{B}_k]$ 为由式(5)为每类电器训练出来的特征矩阵。使用式(7)来预测每类电器用电数据矩阵:

$$\hat{\mathbf{X}}_i = \mathbf{B}_i \hat{\mathbf{A}}_i \tag{7}$$

式中: $\mathbf{B}_i$ 为训练集得到的第 i 类电器的特征矩阵; $\hat{\mathbf{A}}_i$ 为待分解聚合信号找到的最优系数矩阵。综

上,整个求解建模过程是稀疏非负矩阵分解这个研究方向实现能源分解的一个具体过程。本文基于现有工作优化模型,提升能源分解的准确率。

2.1.2 基于 $L_{1/2}$ 范数稀疏约束的非负矩阵分解

在机器学习领域,正则化方法广泛应用于监督学习问题,通常可以简单描述为如下优化问题:

$$w^* = \arg \min_w \sum L(y_i, f(x_i; w)) + \lambda \Omega(w) \tag{8}$$

式中: L 为损失函数用来描述真实值与预测值之间的误差; Ω 为正则项。正则项可以把先验知识加入到模型的学习中去,让学习到的模型具有指定的特性,让经验风险和模型复杂度同时最小。正则化还可以缩小解的范围,加快模型的求解速度,常见的正则项有 L_2 范数、 L_1 范数、 L_0 范数等。其中 L_2 范数可以防止模型过拟合, L_1 范数和 L_0 范数正则项对参数惩罚可使参数稀疏。

L_0 范数、 L_1 范数和 L_2 范数的定义如下:

1) $L_0 = \|\mathbf{x}\|_0 = \#(i) \text{ with } x_i \neq 0$, L_0 表示非零元素的个数。

$$2) L_1 = \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|.$$

$$3) L_2 = \|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}.$$

范数表示的是某个向量空间或者矩阵中每个向量的长度或者大小。因此,范数正则项可以理解为对模型参数解空间添加限制, L_2 范数把参数的解限制在一个圆形的球体内, L_1 范数把参数的解限制在锥形体内, L_0 范数则把参数的解限制在坐标轴上。 L_2 范数正则项使模型的解取圆内的参数,缩小解的范围以防止模型过拟合,提升模型的泛化能力。而 L_1 范数与 L_2 范数不同之处在于 L_1 范数把解约束在锥形体内,等直线最先和解空间相交的地方总是在锥角即坐标轴上,而坐标轴上的点其他维坐标是零。因此, L_1 范数正则项使解中出现零更多,从而达到稀疏约束的目的。

近年来,对 L_1 正则化性能的研究已经取得了很大的突破,然而,在实际应用中 L_1 正则化产生的性能往往小于 L_0 正则化,但是对 L_0 正则化的求解又是一个 NP 难问题。本文通过对范数正则项的分析提出 $L_{1/2}$ 范数正则项来解决 L_1 范数稀疏约束效果不好, L_0 范数不易求解的问题。基于 $L_{1/2}$ 范数稀疏约束的非负矩阵分解能源分解目标函数定义为

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{A}} \quad & \|\mathbf{B} \mathbf{A}_i - \mathbf{X}_i\|_F^2 + \beta \|\mathbf{B}_i\|_{1/2}^{1/2} + \gamma \|\mathbf{A}_i\|_{1/2}^{1/2} \\ \text{s. t.} \quad & \mathbf{B} \geq 0, \mathbf{A} \geq 0 \end{aligned} \tag{9}$$

式中: $\mathbf{X}_i$ 为第 i 类用电器的用电数据矩阵, $i=1,2,3,4,5$ 分别为空调、洗衣机、电炉、冰箱和其他电器的总和; $\mathbf{B}_i$ 为特征矩阵; $\mathbf{A}_i$ 为特征重要程度的权重矩阵; β, γ 为 $L_{1/2}$ 范数正则项的系数,用来控制正则项的强弱。对于 $L_{1/2}$ 正则项定义如下:

$$\|\mathbf{B}\|_{1/2}^{1/2} = \sum_m^M \sum_d^D \mathbf{B}_{m,d}^{\frac{1}{2}} \tag{10}$$

$$\|\mathbf{A}\|_{1/2}^{1/2} = \sum_d^D \sum_l^L \mathbf{A}_{d,l}^{\frac{1}{2}} \tag{11}$$

式(10)和式(11)可以通过如下所示的乘性迭代来求解:

$$\mathbf{A} \leftarrow \mathbf{A} \cdot * \mathbf{B}^T \mathbf{X} \cdot / (\mathbf{B}^T \mathbf{B} \mathbf{A} + \frac{\gamma}{2} \mathbf{A}^{-\frac{1}{2}}) \tag{12}$$

$$\mathbf{B} \leftarrow \mathbf{B} \cdot * \mathbf{X} \mathbf{A}^T \cdot / (\mathbf{B} \mathbf{A} \mathbf{A}^T + \frac{\beta}{2} \mathbf{B}^{-\frac{1}{2}}) \tag{13}$$

式中:符号“ $\cdot *$ ”和“ $\cdot /$ ”表示矩阵的点乘和点除。如果矩阵 $\mathbf{A}$ 或者矩阵 $\mathbf{B}$ 中的元素为 0,可以在 0 元素上加上一个很小的数进行平滑,避免分母为 0。

2.1.3 用户用电模式同质性建模

在社交领域中,同质性表示不同用户之间的相似性^[25]。本文利用家庭用电模式之间的相似性进行建模。

1) 用户用电模式的相似性

利用相似的家庭相似耗能预测,可能得到错误的结论。因此,本文只考虑把相似性作为基础模型分解能源时的一个影响因素,来提高能源分解的准确率。

受此启发,本文通过对不同的数据集进行研究分析发现,总用电量和单类用电器能耗之间也存在着相似性。用电量较高且用电量比较相似的用户在对空调等大功率电器使用上具有一定程度的相似性。用电量较低且用电量比较相似的用户在对照明设备等小功率电器的使用上有着一定的相似性。同样,家庭住宅面积相似的用户在空调设备的用电上存在着相似性。

2) 同质性模型

总用电量相似的家庭在空调等大功率电器的用电量更相似,家庭住宅面积相似的家庭在空调、照明设备等用电器的用电量更相似。本文利用这一先验知识,构建同质性系数,通过添加与同质性系数有关的正则项,来提高模型的分解性能。

首先,本文定义用户 i 和用户 j 之间的同质性系数为 $\varepsilon(i, j)$ 并且满足:

① $\varepsilon(i, j) \in [0, 1]$ 。

② $\varepsilon(i, j) = \varepsilon(j, i)$ 。

③ 同质性系数的值越大,用户之间的相似度就越大,即用户在某一用电器使用量上越相似。

用户在某一用电器用电量的相似性在本文提出的能源分解模型中,可以表示为非负矩阵分解后权重矩阵的相似,因为在特征矩阵确定后权重矩阵决定用户的用电量。因此,在提到用户的用电量相似和权重矩阵相似时表达的意思相同。

对于整个模型可以添加正则项式(14)来约束权重矩阵,从而提高模型的性能。

$$\min \sum_{i=1}^n \sum_{j=1}^n \varepsilon(i, j) \|\mathbf{W}(:, i) - \mathbf{W}(:, j)\|_2^2 \tag{14}$$

式中: $\mathbf{W}(:, i)$ 和 $\mathbf{W}(:, j)$ 分别为矩阵 $\mathbf{W}$ 的第 i 列和第 j 列,表示用户 i 和用户 j 对于某类特定设备的权重向量。对于 2 个不同的用户,某个电器的用电量越相似,在潜在空间中 w_i 和 w_j 的距离也会越近,同时调节它们之间的相似度的同质性系数的值也会越大。

对某一用户 w_i ,其在潜在空间中的同质性正则项可以表示为

$$\sum_{j=1}^n \varepsilon(i, j) \|\mathbf{W}(:, i) - \mathbf{W}(:, j)\|_2^2 \tag{15}$$

式中:同质性系数 $\varepsilon(i, j)$ 控制着用户 w_i 和其他用户之间的关系。对式(15)进行推导,可以得到正则项 $\text{Tr}(\mathbf{W} \mathbf{L} \mathbf{W}^T)$,推导过程如下:

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \varepsilon(i, j) \|\mathbf{W}(:, i) - \mathbf{W}(:, j)\|_2^2 &= \\ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^d \varepsilon(i, j) (\mathbf{W}(k, i) - \mathbf{W}(k, j))^2 &= \\ \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^d \varepsilon(i, j) (\mathbf{W}(k, i))^2 - & \\ \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^d \varepsilon(i, j) (\mathbf{W}(k, i) \times \mathbf{W}(k, j)) &= \\ \sum_{k=1}^d \mathbf{W}(k, :) (\mathbf{D} - \mathbf{Z}) \mathbf{W}^T(k, :) & \end{aligned} \tag{16}$$

式中: $\mathbf{L} = \mathbf{D} - \mathbf{Z}$ 为拉普拉斯矩阵; $\mathbf{D}$ 为对角矩阵,

对角线上元素 $\mathbf{D}(i, i) = \sum_{j=1}^n z(j, i)$; $\mathbf{Z}$ 为由 n 个用户两两之间的同质性系数组成的同质性系数矩阵,定义如下:

$$\mathbf{Z} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1n} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2n} \\ \vdots & \vdots & & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nn} \end{bmatrix}$$

同质性系数 $\varepsilon(i, j)$ 表示相似度,常用的计算相似度的方法有欧氏距离、曼哈度距离、闵可夫斯基距离、余弦距离等。本文使用余弦距离相似度公式来表示用户 i 和用户 j 之间的同质性系数,对于第 k 类用电器,用户 i 和用户 j 之间相似性 $\varepsilon(i, j)$ 定义如下:

$$\varepsilon(i, j) = \frac{\mathbf{U}_i \cdot \mathbf{U}_j}{\|\mathbf{U}_i\| \times \|\mathbf{U}_j\|} \quad (17)$$

式中: $\mathbf{U}_i$ 为用户 i 对第 k 类用电器的用电量向量。

2.1.4 基于用户用电模式相似性建模

由于同质性不能完全代表一个相似家庭的用电数据,本文只考虑把这一性质作为约束项加入基于非负矩阵分解的能源分解模型来改进基础模型,提高模型的性能。从数据出发,本文只考虑加入总用电量同质性和面积同质性。总用电量相似的家庭在单类用电器上存在着相似性,面积相似的家庭在空调的用电量上更相似。首先,计算总用电量同质性系数矩阵,对每类电器单独计算一个用户同质性系数矩阵。对于智能电网用户间的同质性系数通过用户的用电数据计算,训练数据和预测数据采用不同的计算方法。对于训练数据集,因为用户的单类用电数据是已知的,很容易通过余弦相似度公式得到同质性系数矩阵。而对于测试数据集,单类电器的用电量是未知的,此时采用近似计算方法来计算待预测家庭中每类电器的同质性系数,在训练数据集中找到总用电量最接近的 3 个用户,取这 3 个用户对应电器同质性系数的均值来计算。然后,计算面积同质性系数矩阵。对于面积同质性系数,采用相同的计算方法。最后,针对不同类型的电器,本文通过一个系数矩阵来控制同质性正则项的强弱。最终生成基于同质性约束的非负矩阵分解能源分解模型如下:

$$\begin{aligned} \min_{\mathbf{B}_i, \mathbf{A}_i} & \sum_{i=1}^k \|\mathbf{B}_i \mathbf{A}_i - \mathbf{X}_i\|_F^2 + \sum_{i=1}^k \mathbf{E}_{S_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{S_i} \mathbf{A}_i^T) + \\ & \sum_{i=1}^k \mathbf{E}_{f_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{f_i} \mathbf{A}_i^T) \\ \text{s. t. } & \mathbf{B}_i \geq 0, \mathbf{A}_i \geq 0 \quad i = 1, 2, \dots, k \end{aligned} \quad (18)$$

式中: $\text{Tr}(\mathbf{A}_i \mathbf{L}_{S_i} \mathbf{A}_i^T)$ 和 $\text{Tr}(\mathbf{A}_i \mathbf{L}_{f_i} \mathbf{A}_i^T)$ 分别为总用电量同质性正则项和面积同质性正则项; $\mathbf{E}_S$ 和 $\mathbf{E}_f$ 分别为总用电量和面积同质性惩罚项的权重,用于控制同质性正则项强弱; $\mathbf{L}_S$ 和 $\mathbf{L}_f$ 分别为总用电量拉普拉斯矩阵和面积拉普拉斯矩阵。

2.1.5 能源分解方法建模

在本节中,将提出本文最终的能源分解方法,再进一步优化,并给出求解过程。

1) 基于 $L_{1/2}$ 范数稀疏约束和同质性约束的能源分解方法,建立的能源分解训练模型如下:

$$\begin{aligned} \min_{\mathbf{B}_i, \mathbf{A}_i} & \sum_{i=1}^k \|\mathbf{B}_i \mathbf{A}_i - \mathbf{X}_i\|_F^2 + \beta \sum_{i=1}^k \|\mathbf{B}_i\|_{1/2}^{1/2} + \\ & \gamma \sum_{i=1}^k \|\mathbf{A}_i\|_{1/2}^{1/2} + \sum_{i=1}^k \mathbf{E}_{S_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{S_i} \mathbf{A}_i^T) + \end{aligned}$$

$$\sum_{i=1}^k \mathbf{E}_{f_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{f_i} \mathbf{A}_i^T)$$

$$\text{s. t. } \mathbf{B}_i \geq 0, \mathbf{A}_i \geq 0 \quad i = 1, 2, \dots, k \quad (19)$$

式中: β, γ 为 $L_{1/2}$ 稀疏约束正则项系数; $\mathbf{E}_S$ 和 $\mathbf{E}_f$ 为同质性正则项的系数。

$$\begin{aligned} \hat{\mathbf{A}}_{1:k} &= \arg \min_{\mathbf{A}_{1:k} \geq 0} \left\| \bar{\mathbf{X}} - [\mathbf{B}_1 \quad \dots \quad \mathbf{B}_k] \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_k \end{bmatrix} \right\|_F^2 + \\ & \beta \sum_{i=1}^k \|\mathbf{B}_i\|_{1/2}^{1/2} + \gamma \sum_{i=1}^k \|\mathbf{A}_i\|_{1/2}^{1/2} + \\ & \sum_{i=1}^k \mathbf{E}_{S_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{S_i} \mathbf{A}_i^T) + \\ & \sum_{i=1}^k \mathbf{E}_{f_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{f_i} \mathbf{A}_i^T) = \\ & \arg \min_{\mathbf{A}_{1:k} \geq 0} F(\bar{\mathbf{X}}, \mathbf{A}_{1:k}, \mathbf{B}_{1:k}) \end{aligned} \quad (20)$$

$$\hat{\mathbf{X}}_i = \mathbf{B}_i \hat{\mathbf{A}}_i \quad (21)$$

对于训练数据集,使用目标函数式(19)求解得到每类电器最优的特征矩阵 $\mathbf{B}_i$ 。对于测试集,用式(20)为每类电器找到最优的权重矩阵 $\mathbf{A}_i$,最终用式(21)预测每类电器的用电量。

2) 模型优化与求解。如果训练过程和分解的过程能够同步进行,那么在求解过程中这 2 个结果会相互影响,对预测结果会有正向促进作用。因此,本文继续改进模型得到最终能源分解模型如下:

$$\begin{aligned} \min_{\mathbf{W}_i, \mathbf{B}_i, \mathbf{A}_i} & \left\| \sum_{i=1}^k \mathbf{B}_i \mathbf{W}_i - \mathbf{G} \right\|_F^2 + \sum_{i=1}^k \|\mathbf{B}_i \mathbf{A}_i - \mathbf{X}_i\|_F^2 + \\ & \lambda \sum_{i=1}^k \|\mathbf{W}_i \mathbf{H} - \mathbf{A}_i\|_F^2 + \alpha \sum_{i=1}^k \|\mathbf{W}_i\|_{1/2}^{1/2} + \\ & \beta \sum_{i=1}^k \|\mathbf{B}_i\|_{1/2}^{1/2} + \gamma \sum_{i=1}^k \|\mathbf{A}_i\|_{1/2}^{1/2} + \\ & \sum_{i=1}^k \mathbf{E}_{S_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{S_i} \mathbf{W}_i^T) + \sum_{i=1}^k \mathbf{E}_{f_i} \odot \text{Tr}(\mathbf{A}_i \mathbf{L}_{f_i} \mathbf{W}_i^T) \\ \text{s. t. } & \mathbf{B}_i \geq 0, \mathbf{A}_i \geq 0, \mathbf{W}_i \geq 0 \quad i = 1, 2, \dots, k \end{aligned} \quad (22)$$

式中: $\mathbf{W}, \mathbf{B}$ 和 $\mathbf{A}$ 为待求矩阵; 矩阵 $\mathbf{G}$ 为训练数据和测试数据联结的矩阵, $\mathbf{G}$ 的前 d 列表达的数据与矩阵 $\mathbf{X}$ 所表达的是同一批训练数据。矩阵 $\mathbf{G}$ 和 $\mathbf{X}$ 重合部分的数据权重应该尽可能一致,即 $\mathbf{W}(1:d)$ 与 $\mathbf{A}$ 尽可能一致。因此,在最终模型中对权重矩阵 $\mathbf{A}$ 的同质性约束可以包含在权重矩阵 $\mathbf{W}$ 内。为了实现 $\mathbf{W}(1:d)$ 与 $\mathbf{A}$ 尽可能一致,本文在目标函数中加入了一致性正则项 $\sum_{i=1}^k \|\mathbf{W}_i \mathbf{H} - \mathbf{A}_i\|_F^2$, 其中 $\mathbf{H}$ 为扩展的对角矩阵, $\mathbf{W}\mathbf{H}$ 可实现取 $\mathbf{W}$ 矩阵的前 d 列,从而实现加入此

正则项后使矩阵 $\mathbf{W}$ 和矩阵 $\mathbf{A}$ 尽可能一致。

在上述内容中,已经得到了本文提出的最终的能源分解模型,如式(22)所示,但是该公式无法直接给出 $\mathbf{W}$ 、 $\mathbf{B}$ 、 $\mathbf{A}$ 的解,本文将采用乘性迭代法交替更新 $\mathbf{W}$ 、 $\mathbf{B}$ 、 $\mathbf{A}$ 来找到其最优解,具体过程本文不再赘述,迭代规则如下:

$$\mathbf{W}_i(j, n) \leftarrow \mathbf{W}_i(j, n) \cdot$$

$$\sqrt{\frac{[\mathbf{B}_i^T \mathbf{G} + \lambda \mathbf{A}_i \mathbf{H}^T + \mathbf{U}_i]_{j, n}}{\left[\sum_{l=1}^k \mathbf{B}_l^T \mathbf{B}_l \mathbf{W}_l + \lambda \mathbf{W}_i \mathbf{H} \mathbf{H}^T + \mathbf{O}_i - \mathbf{V}_i \right]_{j, n}}} \quad (23)$$

$$\mathbf{O}_i = \frac{\alpha}{2} (\mathbf{W}_i + \eta)^{-\frac{1}{2}} \quad (24)$$

$$\mathbf{U}_i = \mathbf{E}_{S_i} \mathbf{W}_i \mathbf{Z}_{S_i} + \mathbf{E}_{f_i} \mathbf{W}_i \mathbf{Z}_{f_i} \quad (25)$$

$$\mathbf{U}_i = \mathbf{E}_{S_i} \mathbf{W}_i \mathbf{Z}_{S_i} + \mathbf{E}_{f_i} \mathbf{W}_i \mathbf{D}_{f_i} \quad (26)$$

$$\mathbf{B}_i(j, n) \leftarrow \mathbf{B}_i(j, n) \cdot$$

$$\sqrt{\frac{[\mathbf{G} \mathbf{W}_i^T + \mathbf{X}_i \mathbf{A}_i^T]_{j, n}}{\left[\sum_{l=1}^k \mathbf{B}_l \mathbf{W}_l \mathbf{W}_l^T + \mathbf{B}_i \mathbf{A}_i \mathbf{A}_i^T + \frac{\beta}{2} (\mathbf{B}_i + \eta)^{-\frac{1}{2}} \right]_{j, n}}} \quad (27)$$

$$\mathbf{A}_i(j, n) \leftarrow \mathbf{A}_i(j, n) \cdot$$

$$\sqrt{\frac{[\mathbf{B}_i^T \mathbf{X}_i + \lambda \mathbf{W}_i \mathbf{H}]_{j, n}}{\left[\mathbf{B}_i^T \mathbf{B}_i \mathbf{A}_i + \lambda \mathbf{A}_i + \frac{\gamma}{2} (\mathbf{A}_i + \eta)^{-\frac{1}{2}} \right]_{j, n}}} \quad (28)$$

通过以上公式迭代计算,最终可以得到基函数矩阵 $\mathbf{B}_i$ 和系数矩阵 $\mathbf{W}_i$,使用式(29)对测试数据进行分解,重构矩阵 $\hat{\mathbf{X}}_i$ 。

$$\text{New} \hat{\mathbf{X}}_i = \mathbf{B}_i \mathbf{W}_i \quad (29)$$

2.2 基于能源分解的用户用电行为聚类分析方法

2.2.1 基于改进 K-Means 聚类算法的用户聚类模型

选择 K-Means 聚类算法来辅助验证本文提出的观点。首先,介绍传统的基于用电行为的 K-Means 聚类算法。

假设数据集 $\mathbf{I} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \in \mathbf{R}^T$, $\mathbf{x}_n$ 是 T 维的向量,表示用户 n 在 T 段时间内的总用电量,具体算法描述如下:

步骤 1 随机初始化 k 个聚类中心 $u_1, u_2, \dots, u_k \in \mathbf{R}^T$ 。

步骤 2 针对每个样本点,计算其与每一个类中心的距离,离哪个类中心近就划分到哪个类中,这样完成一次聚类。

步骤 3 用类均值更新聚类中心。

步骤 4 判断新计算出的类中心和原来的类中心之间的距离是否小于设定的阈值(表示重新

计算的类中心位置变化不大,趋于稳定),如果小于阈值,算法终止,否则返回步骤 2。K-Means 聚类算法中通常选择欧氏距离来计算点到类中心的距离,对于样本点到类中心的距离计算如下:

$$d(\mathbf{x}_i, u_j) = \sqrt{\sum_{t=1}^T (x_{it} - u_{jt})^2} \quad (30)$$

式中: x_{it} 为用户 i 在 t 时刻的用电量; u_{jt} 为聚类中心 j 在 t 时刻的数据。

为了避免总用电特征对用户聚类结果不够准确及不能反映用户的详细用电行为的问题。本文提出用能源分解后的单类用电器的用电特征来代替总用电特征,计算用户间的距离提升聚类效果,改进传统的 K-Means 聚类算法。

具体地,对于一个用户 i ,传统的方法是用一个向量来表示,该向量中的值是用户的总用电特征,现在我们用一个矩阵来表示用户 i ,矩阵中每一列是一个单独用电器的用电向量, K 列就有 K 个用电器。在计算距离之前用 Z-Score 消除不同类电器量级不同带来的影响,改进后的算法的距离计算为

$$d(\mathbf{X}_i, \mathbf{U}_j) = \sqrt{\sum_{k=1}^K \sum_{t=1}^T (X_{ikt} - U_{jkt})^2} \quad (31)$$

式中: $\mathbf{X}_i$ 为用户 i 的特征矩阵; $\mathbf{U}_j$ 为聚类中心 j 的特征矩阵; X_{ikt} 为用户 i 用电器 k 在 t 时刻的用电量; U_{jkt} 为聚类中心 j 对于用电器 K 在 t 时刻的类中心数据。

2.2.2 初始聚类中心和 K 值的选择

K-Means 聚类算法的第一步也是最重要的一步,即选择聚类数目 K 值和选择初始聚类中心。 K 值决定聚类结果的数目, K 值不同聚类结果也不同,聚类结果对 K 值的依赖性在一定程度上影响着聚类分析。而聚类中心的选择影响聚类结果的稳定性,并且容易使算法陷入局部最优解。基本的初始聚类中心的选择是随机的,会使目标结果偏离最优结果,从而陷入局部最优解。因此,合适的 K 值和初始聚类中心对于智能电网用户的聚类结果至关重要。

本文采用最大距离法选择初始类中心。选取的初始类中心对象之间的距离应该尽可能离得远。首先,随机选取一个样本作为第一个类中心,然后,选择与第一个类中心最远的样本作为第二个类中心,接着用同样的方法选择其他的类中心。

本文选择聚类有效性指标来计算合适的聚类数目。常用的评测指标有手肘法、轮廓系数(silhouette coefficient, SC)、戴维斯-布尔丁指数(Davies-Bouldin index, DBI)、卡林斯基-哈拉巴斯指数(Calinski-Harabaz index, CHI,也被称为方差比标

准)及权变矩阵。本文选择手肘法、DBI 这 2 种内部有效性指标来选择最优的 K 值。这 2 种指标的计算方式如下:

1) 手肘法。手肘法的核心指标是误差平方和(sum of the squared errors,SSE)计算如下:

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2 \quad (32)$$

式中: C_i 为第 i 个簇; m_i 为 C_i 的簇质心; p 为 C_i 中的样本点;SSE 为所有样本的聚类误差,代表着聚类效果的好坏。

手肘法的核心思想是:随着聚类数目 K 值的增多,每个簇的聚合程度也会逐渐增加,而 SSE 则会逐渐减小;并且当 K 值在小于最优聚类数目范围内增加时,SSE 的下降幅度会比较大,当 K 值增加到最优聚类数目时,SSE 的下降幅度则会骤减,然后随着 K 值的继续增大 SSE 会慢慢趋于平缓。SSE 和 K 值的变化关系类似于手肘的形状,并且在肘部取得最优的 K 值,因此称为手肘法。

2) DBI。DBI 有效性指标定义如下:

$$DBI = \frac{1}{N} \sum_{i=1}^N \max_{j \neq i} \left(\frac{S_i + S_j}{\|\omega_i - \omega_j\|_2} \right) \quad (33)$$

式中: S_i 为第 i 类内的样本点到类中心的平均距离,代表了第 i 类的分散程度; $S_i + S_j$ 为簇 i 和簇 j 的分散程度; $\|\omega_i - \omega_j\|_2$ 为簇 i 和簇 j 质心之间的距离,用簇 i 和簇 j 的分散度和距离之商来表示这 2 个簇 i 和簇 j 的相似度。对于每一个簇 i ,找到最相似的簇 $j(i \neq j)$,即最差的聚类情况。最终计算所有簇的最大平均相似度作为 DBI 指标,DBI 越小表示聚类效果越好。

3 实验结果与分析

3.1 基于 $L_{1/2}$ 范数和同质性约束的能源分解方法实验

3.1.1 不同方法对比实验结果分析

对基本的非负矩阵分解的能源分解方法(NMF)、 L_1 范数约束的稀疏编码方法(NMF + L_1)、基于 $L_{1/2}$ 正则项约束的能源分解方法(NMF + $L_{1/2}$)、本文方法 4 种方法做实验。使用同一数据集,在参数最优的情况下经过多次重复实验,性能取均值的结果如表 1 和图 2 所示。

从表 1 中可以看出,在基本的非负矩阵分解模型上加入约束项之后,算法的性能都有很大的提升,并且模型的训练时间没有太大的波动。加入 L_1 范数稀疏约束的模型相比于 $L_{1/2}$ 范数的稀疏约束的模型在训练集上效果要好,但是在测试集

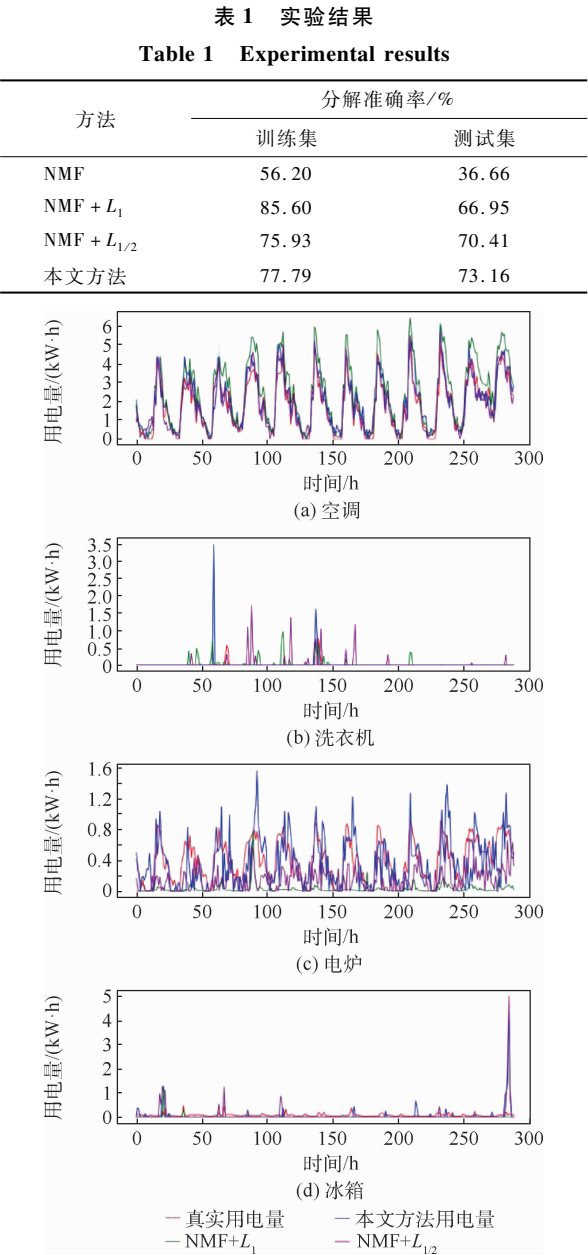


图 2 真实结果和分解结果对比

Fig. 2 Comparison of real results and decomposition results 上没有 $L_{1/2}$ 范数的稀疏约束效果好,总体上 $L_{1/2}$ 范数的稀疏约束效果更好。本文所提出的方法,在 NMF 的基础上加入 $L_{1/2}$ 范数稀疏约束项、同质性约束项和一致性约束项,模型的性能在训练集和测试集上都取得了相对较好的结果。

如图 2 所示,对于能源分解后空调的用电量 3 种对比方法都能很好的拟合,但是洗衣机、电炉、冰箱这 3 类电器本文提出的方法拟合效果更好。

3.1.2 同质性对实验结果的影响

1) 总用电量同质性
总用电量同质性对实验结果的影响如图 3 所示,其中 x 轴为小功率电器参数 E_1 , y 轴为大功率电器参数 E_h , z 轴表示分解准确率(A_{cc})。分解准确率随着参数 E_h 和 E_1 的变化呈现明显的规律性。

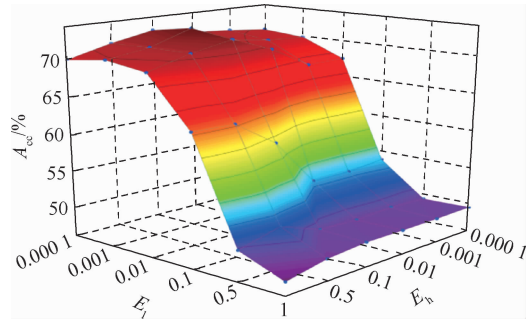


图3 不同参数对比

Fig. 3 Contrast diagram of different parameters

参数 E_h 变化时对分解准确率影响较大,当参数的取值大于0.01分解性能急剧下降,当参数小于0.01时分解准确率相对稳定且较高,在0.001附近取得最优值。而分解准确率在参数 E_l 的影响下并没有呈现这种规律,参数 E_l 对分解准确率影响相对较小,随着 E_l 的增大模型的性能先增大后减小,在0.1附近取得最优值。

2) 面积同质性

面积同质性对实验结果的影响如表2所示。可以看出,空调的同质性参数 E_t 取不同值时,模型的分解性能呈规律性变化,参数小于0.1时模型的性能相对较好,在0.001附近取得最优。

为了清晰地显示面积对空调分解性能的影响,随机选取一个用户分别画出空调2周内和1天内真实电量和预测电量之间的变化趋势,如图4所示,2条线变化趋势相同并且基本重合,预测值接近真实值。

表2 参数 E_t 对比结果

Table 2 Parameter E_t contrast results						
E_t	0.000 1	0.001	0.01	0.1	0.5	1
$A_{cc}/\%$	69.42	69.72	66.74	59.22	58.74	58.77

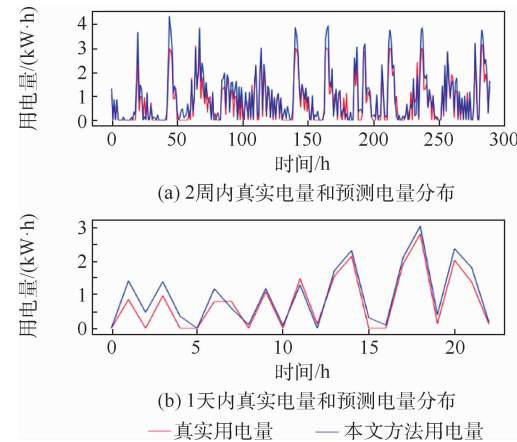


图4 2周内和1天内真实电量和预测电量分布

Fig.4 Predicted distribution of real power consumption in two week and predicted distribution of real power consumption in one day

3.2 基于能源分解的用户用电行为聚类分析方法实验

3.2.1 最优聚类数目K的选择实验

1) 传统的 K-Means 聚类算法 K 值确定

基于传统的 K-Means 聚类算法使用的是用户总用电特征进行聚类分析,即进行能源分解之前智能电表直接采集到的总用电特征数据集进行实验,采用欧氏距离计算用户之间的距离。聚类数目 K 取 $[2,10]$,DBI 和 SSE 这2种有效性指标在不同聚类数目 K 下的结果分别如图5和图6所示,图中横轴表示聚类数目 K ,纵轴表示有效性指标。

从图5中可以看出,当 $K=4$ 时到达SSE的变化曲线的肘部,因此在SSE有效性指标下最佳聚类数目为4。从图6中可以看出,DBI最小值为3,这表示在DBI指标下最佳聚类数目为3。但是,值得注意的是当 K 取4时DBI的值还非常大,而 $K=3$ 或 $K=4$ 时SSE差别并不大,因此本文选择 $K=3$ 为传统的 K-Means 聚类算法的最佳聚类数目。

2) 改进的 K-Means 聚类算法 K 值确定

本文提出的聚类算法使用的是用户单类用电

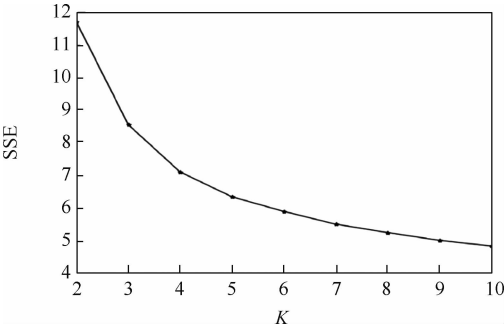


图5 传统的 K-Means 聚类算法有效性指标 SSE 评测结果

Fig.5 Traditional K-Means clustering algorithm evaluation results of effectiveness index of traditional clustering SSE

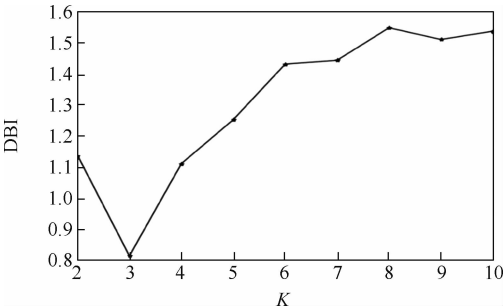


图6 传统的 K-Means 聚类算法有效性指标 DBI 评测结果

Fig.6 Traditional K-Means clustering algorithm evaluation results of effectiveness index of traditional clustering DBI

器的用电特征进行聚类分析的,即能源分解之后得到的数据,同样采用欧氏距离计算用户之间的距离,使用 SSE 和 DBI 这 2 种有效性指标选择最优的聚类数目 K 。随着 K 值的变化,SSE 和 DBI 的变化曲线如图 7 和图 8 所示,图中横轴表示聚类数目 K ,纵轴表示有效性指标。

从图 7 中可以看出,当 $K=4$ 时 SSE 曲线到达肘部,在 SSE 有效性指标下最佳聚类数目 K 为 4。

从图 8 中可以看出,DBI 在 $K=2$ 时取得最小值,即最佳聚类数目 K 为 2,但是在 $K=2$ 时 SSE 还非常大,这不具有合理性。因此,本文退而求其次选择第二小 DBI 值处的聚类数目 K ,即 $K=4$ 。综合分析,改进的 K-Means 聚类算法最佳聚类数目 K 应为 4。

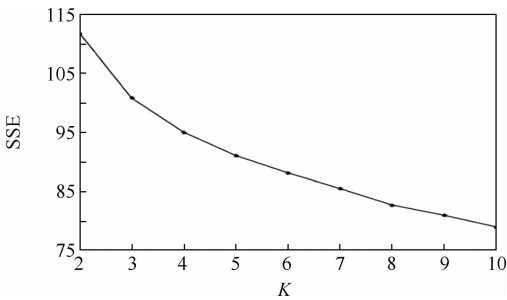


图 7 改进的 K-Means 聚类算法有效性指标 SSE 评测结果

Fig. 7 Improved K-Means clustering algorithm evaluation result of effectiveness index of improved clustering SSE

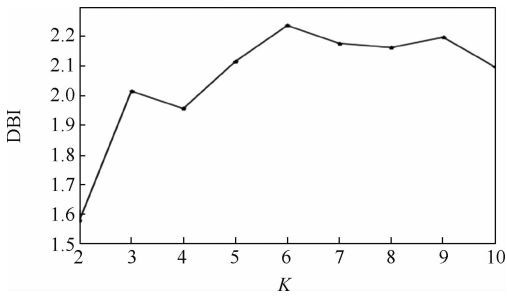


图 8 改进的 K-Means 聚类算法有效性指标 DBI 评测结果

Fig. 8 Improved K-Means clustering algorithm evaluation result of effectiveness index of improved clustering DBI

3.2.2 不同方法对比实验

1) 传统 K-Means 聚类算法实验分析

当 $K=3$ 时,传统的 K-Means 聚类算法的聚类结果如图 9 所示,红绿蓝 3 条线分别代表 3 个类的类中心。图中:cluster0_139 表示类 0,clusterl_91 表示类 1,cluster2_34 表示类 2。从图中可以看出,蓝线代表的类中心变化幅度比较小,几乎趋于平稳,红线和绿线变化趋势相同,都是在 5:00 之

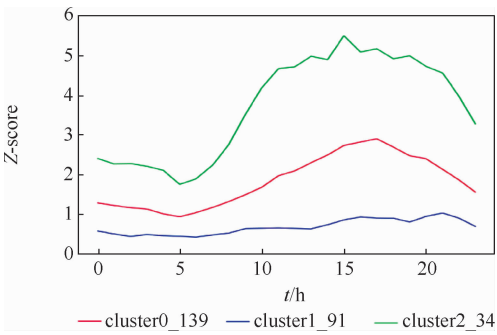


图 9 总用电量用户聚类结果

Fig. 9 Clustering results for users with total electricity consumption

后用电量慢慢上升,19:00用电量开始下降,但是变化幅度不同,绿线代表的用户变化幅度比较大,用电量比较多。对总用电数据聚类能反映用户的总体用电趋势,而每个分电器的用电特点无从得知,用电峰值也不突出,这在一定程度上会影响聚类结果的准确性。

2) 改进的 K-Means 聚类算法实验分析

当 $K=4$ 时,对用户的单独电器进行聚类算法的聚类结果如图 10 所示,图中每一行表示一类用户,从上到下依次为第 1 类、第 2 类、第 3 类、第 4 类用户。图 10(a) ~ (d) 分别代表 4 种用电器(空调、洗衣机、电炉、冰箱)的日用电特点。从图中可以看出,空调和电炉的用电趋势比较相似;洗衣机的日用电量,每个家庭使用时段不同,呈现出不同的特点;冰箱的日用电量,也可以通过电量分析,大致了解到每类用户冰箱的使用情况。实验中的 4 类用户中,后面 3 类用电都很平稳,第 1 类变化波动比较大并且相对用电量比较高,可能第 1 类用户更注重家庭饮食品质,冰箱存放食物更多。第 1 类和第 4 类用户洗衣机日用电量波动比较大,在上午和下午出现 2 次波峰,第 2、3 类用户无明显波动,这一天可能没有使用洗衣机。从图中也可以看出,第 3 类用户的 4 种用电器电量基本上没有浮动,都处于关闭状态,可以认为这类用户在这一天没有使用电器。

3) 对比分析

从 2 种对比方法的实验结果中可以看到,本文提出的聚类方法更能准确分析用户的行为,识别每类用户中单独电器的用电特点。针对每类用户的单类用电器的用电特点做出更有数据支撑、更合理的电网规划、需求响应和信息推荐等决策。由于数据集的局限本文只分析了 4 类电器,如果能够获得更多单类电器的用电数据,聚类结果将会有更大的应用价值。

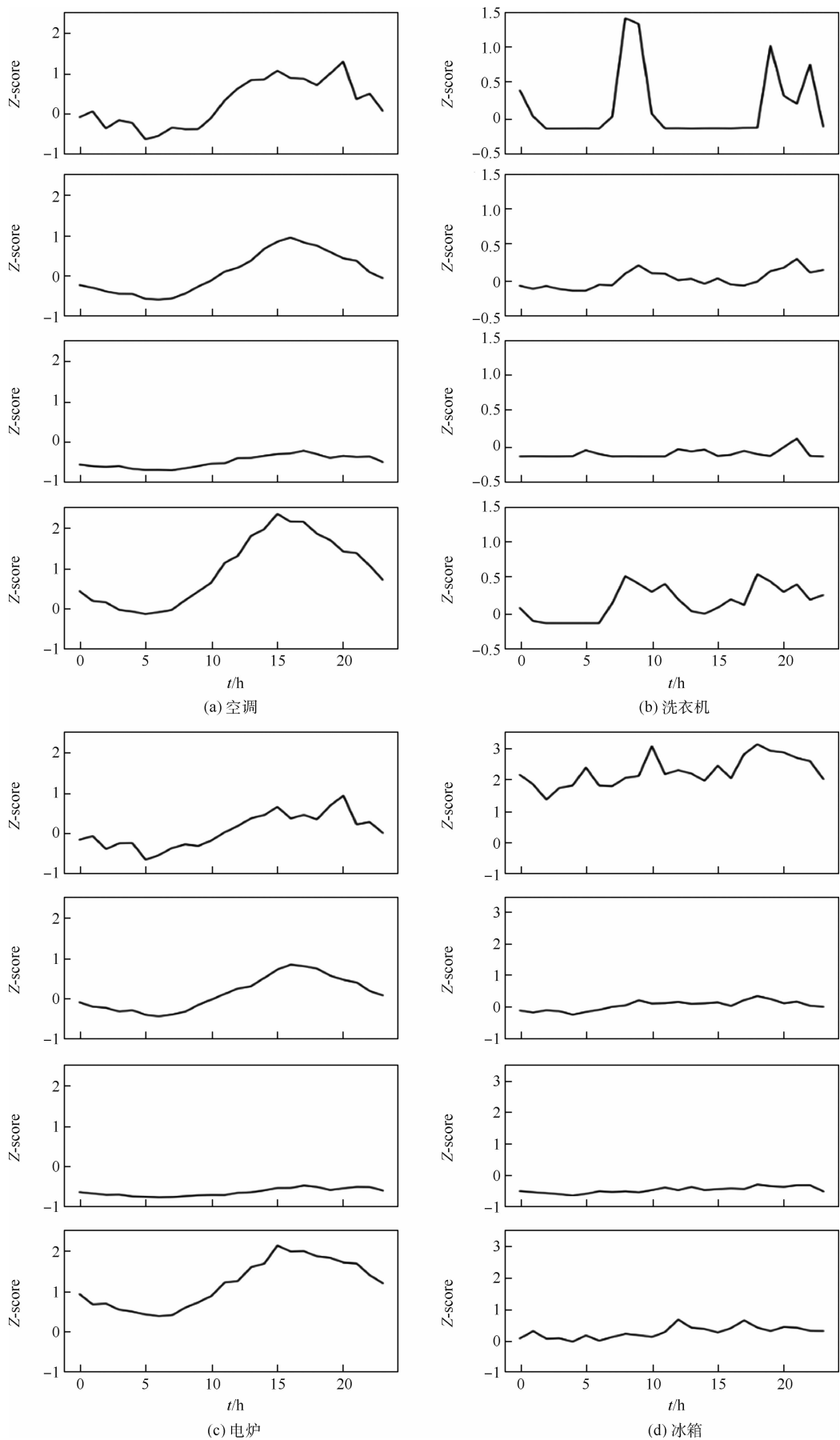


图 10 单类用电器用户聚类结果

Fig. 10 Clustering result for users with single-type electrical appliances

4 结 论

随着智能电网和大数据技术的快速发展,大量用户的用电数据被记录和存储,如何合理、有效地利用这些数据挖掘出数据中潜在的价值越来越受到关注。能源分解和用户用电行为分析作为智能电网应用之一,能够用于节约能源、故障检测、电网规划、需求响应和信息推荐等。因此,对用户的电能分解并对其用电行为进行聚类分析具有重要价值,而如何提高能源分解的准确率和用户用电行为聚类效果变得极为重要。基于此,本文提出了一种基于能源分解的用户用电行为模式分析方法。采用非负矩阵分解的能源分解方法(NMF)结合 $L_{1/2}$ 范数的稀疏约束、同质性约束项及一致性约束项方法训练数据,得到的模型在训练集和测试集上都具有很好的效果,克服了传统采用 NMF + L_1 方法得到的模型在测试集上效果不好的问题及克服了采用 NMF + $L_{1/2}$ 方法精度不高的问题。本文在传统方法 NMF + $L_{1/2}$ 上加入同质性约束项及一致性约束项训练的模型,一定程度上提高了模型在训练集和测试集上的性能。

参考文献 (References)

[1] DIECJMANN J, COOPERMAN A, BRODRICK J. Efficiency energy. Buildings energy data book [M]. Washington, D. C. : US Department of Energy, 2009.

[2] 辛苗苗,张延迟,解大. 基于电力大数据的用户用电行为分析研究综述[J]. 电气自动化, 2019, 41(1) : 1-4.

XIN M M, ZHANG Y C, XIE D. Summary of researches on consumer behavior analysis based on big power data [J]. Electrical Automation, 2019, 41(1) : 1-4 (in Chinese).

[3] HART G W. Nonintrusive appliance load monitoring [J]. Proceedings of the IEEE, 1992, 80(12) : 1870-1891.

[4] LAUGHMAN C, LEE K, COX R, et al. Power signature analysis [J]. IEEE Power and Energy Magazine, 2003, 1(2) : 56-63.

[5] SHAW S R, ABLER C B, LEPARD R F, et al. Instrumentation for high performance nonintrusive electrical load monitoring [J]. Journal of Solar Energy Engineering, 1998, 120(3) : 224-229.

[6] BERGES M, GOLDMAN E, MATTHEWS H S, et al. Learning systems for electric consumption of buildings [C] // International Workshop on Computing in Civil Engineering 2009. Reston: American Society of Civil Engineers, 2009 : 1-10.

[7] PATEL S N, ROBERTSON T, KIENTZ J A, et al. At the flick of a switch: Detecting and classifying unique electrical events on the residential power line (nominated for the best paper award) [C] // International Conference on Ubiquitous Computing. Berlin: Springer, 2007 : 271-288.

[8] ZEIFMAN M, ROTH K. Nonintrusive appliance load monitoring: Review and outlook [J]. IEEE Transactions on Consumer Electronics, 2011, 57(1) : 76-84.

[9] KOLTER J, BATRA S, NG A. Energy disaggregation via discriminative sparse coding [J]. Advances in Neural Information Processing Systems, 2010, 23 : 1153-1161.

[10] KIM H, MARWAH M, ARLITT M, et al. Unsupervised disaggregation of low frequency power measurements [C] // Proceedings of the 2011 SIAM International Conference on Data Mining, 2011 : 747-758.

[11] PARSON O, GHOSH S, WEAL M, et al. An unsupervised training method for non-intrusive appliance load monitoring [J]. Artificial Intelligence, 2014, 217 : 1-19.

[12] KOLTER J Z, JAAKKOLA T. Approximate inference in additive factorial hmms with application to energy disaggregation [C] // Artificial Intelligence and Statistics, 2012 : 1472-1482.

[13] KELLY J, KNOTTENBELT W. Neural NILM: Deep neural networks applied to energy disaggregation [C] // Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments. New York: ACM, 2015 : 55-64.

[14] SIROJAN T, PHUNG B T, AMBIKAIRAJAH E. Deep neural network based energy disaggregation [C] // 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE). Piscataway: IEEE Press, 2018 : 73-77.

[15] GAO Y, SCHAY A, HOU D Q, et al. Home appliance energy disaggregation using low frequency data and machine learning classifiers [C] // 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE Press, 2017 : 76-83.

[16] BATRA N, SINGH A, WHITEHOUSE K. Gemello: Creating a detailed energy breakdown from just the monthly electricity bill [C] // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016 : 431-440.

[17] 李培强,李欣然,陈辉华,等. 基于模糊聚类的电力负荷特性的分类与综合 [J]. 中国电机工程学报, 2005, 25(24) : 73-78.

LI P Q, LI X R, CHEN H H, et al. The characteristics classification and synthesis of power load based on fuzzy clustering [J]. Proceedings of the CSEE, 2005, 25(24) : 73-78 (in Chinese).

[18] 王璨,冯勤超. 基于价值评价的电力用户分类研究 [J]. 价值工程, 2009, 28(5) : 64-67.

WANG C, FENG Q C. The research of power customers classification based on value assessment [J]. Value Engineering, 2009, 28(5) : 64-67 (in Chinese).

[19] 李欣然,姜学皎,钱军,等. 基于用户日负荷曲线的用电行业分类与综合方法 [J]. 电力系统自动化, 2010, 34(10) : 56-61.

LI X R, JIANG X J, QIAN J, et al. A classifying and synthesizing method of power consumer industry based on the daily load profile [J]. Automation of Electric Power Systems, 2010, 34(10) : 56-61 (in Chinese).

[20] ZHONG C L, SHAO J, ZHENG F, et al. Research on electricity consumption behavior of electric power users based on tag technology and clustering algorithm [C] // 2018 5th International Conference on Information Science and Control Engineering

(ICISCE). Piscataway:IEEE Press,2018:459-462.

[21] NORDAHL C,BOEVA V,GRAHN H,et al. Profiling of house-
hold residents' electricity consumption behavior using cluster-
ing analysis[C] // International Conference on Computational
Science. Berlin:Springer,2019:779-786.

[22] LEE D D,SEUNG H S. Learning the parts of objects by non-
negative matrix factorization [J]. Nature, 1999, 401 (6755) :
788-791.

[23] OLSHAUSEN B A, FIELD D J. Emergence of simple-cell re-
ceptive field properties by learning a sparse code for natural im-
ages[J]. Nature,1996,381 (6583) :607-609.

[24] RAHIMPOUR A,QI H R,FUGATE D,et al. Non-intrusive en-
ergy disaggregation using non-negative matrix factorization with
sum-to- k constraint[J]. IEEE Transactions on Power Systems,
2017,32 (6) :4430-4441.

[25] TANG J,GAO H,HU X,et al. Exploiting homophily effect for
trust prediction[C] // Proceedings of the 6th ACM Interhational
Conference on Web Search and Data Mining. New York:ACM,
2013:53-62.

User electricity consumption behavior mode analysis

based on energy decomposition

LU Ruirui, YU Haiyang, YANG Zhen^{*}, LAI Yingxu, YANG Shisong, ZHOU Ming

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: With the popularization of smart grids and the development of big data technology, more and more attention has been paid to the analysis of users' electricity consumption behavior through electricity consumption data. The existing energy decomposition methods cannot meet the high requirements for resolution and decomposition accuracy in practical applications, and the cluster analysis method is too rough and does not fully show the electricity consumption characteristics of each type of electrical appliances. In view of this, this paper proposes an analysis method of users' electricity consumption behavior based on energy decomposition. Based on the discriminative sparse coding algorithm model, firstly, to solve the problem that the regular term of L_0 is not easy to solve and the effect of the sparse constraint of L_1 regular term is not ideal, we propose to use the sparse constraint of $L_{1/2}$ regular term to perform energy decomposition, and add the homogeneity between users as a regular term to the basic model to modify the performance of the model. Secondly, based on the results of energy decomposition, we use the electricity consumption characteristics of a user's single-type electrical appliances instead of the total electricity consumption characteristics to refine the analysis of user's electricity consumption behavior, and improve the traditional K-Mean clustering algorithm for experimental verification. The experimental results show that the energy decomposition method based on the sparse constraint of $L_{1/2}$ regular term and the constraint of homogeneity can effectively improve the accuracy of energy decomposition compared with the traditional discriminative sparse coding method. At the same time, the result of cluster analysis of users' electricity consumption behavior based on energy decomposition is also significantly improved.

Keywords: smart grid; energy decomposition; cluster analysis; nonnegative matrix decomposition; homogeneous relationship

Received: 2020-09-27; Accepted: 2020-11-06; Published online: 2021-03-10 10:33

URL: kns.cnki.net/kcms/detail/11.2625.V.20210309.1030.004.html

Foundation items: National Natural Science Foundation of China (61671030); Beijing Great Wall Scholar (CIT&TCD20190308); China Post-doctoral Science Foundation (2019M660377)

* Corresponding author. E-mail: yangzhen@bjut.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0398

面向智能电网嵌入式设备的网络威胁动态评估方法

吕卓^{1,*}, 郭志民¹, 陈岑¹, 莫坚松¹, 常朝稳²

(1. 国网河南省电力公司电力科学研究院, 郑州 450052; 2. 中国人民解放军战略支援部队信息工程大学, 郑州 450004)

摘 要: 针对智能电网嵌入式设备由于计算、存储资源有限而造成的对网络攻击行为应对不足, 安全评估手段薄弱等问题, 提出了面向智能电网嵌入式设备的网络攻击行为动态评估方法。使用安全控制模块对实际嵌入式设备通信数据流进行解析与判别, 利用组件动态可信度量分析方法在嵌入式系统模拟机中对攻击行为安全影响进行安全检测评估, 通过对平台配置属性、平台运行属性及用户认证属性 3 个方面属性进行全过程动态综合度量, 得出最终网络攻击行为安全评估结果。通过在配电自动化及用电信息采集系统真实环境下进行测试, 针对嵌入式设备常见的攻击行为, 检测方法的准确率能够达到 90% 以上, 具备较好的安全评估精度, 与此同时实现了自身安全性的有效提升。

关 键 词: 动态评估; 属性度量; 安全检测; 网络攻击行为; 动态可信度量

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2022)02-0324-07

近年来, 国内外由于嵌入式设备遭受攻击造成的电力系统安全事故屡有发生, 2015 年 12 月 23 日发生的乌克兰电网遭受黑客攻击导致大面积用户停电事件, 成为世界上首例由于黑客入侵而导致的大面积停电事故, 如何保障电网控制系统的安全稳定运行已经成为了各国保障国家安全的重要研究课题。

相对于传统 PC 设备, 智能电网设备关注更优的硬件适配、更少的资源占用及更可靠的系统运行等问题, 而对其自身信息安全缺少足够的防护, 攻击者很容易通过各种安全漏洞侵入到智能电网设备中实施干扰、监视甚至远程控制^[1-3]。由于嵌入式设备自身计算及存储资源有限, 面对各类网络攻击行为时显的力不从心^[4-6], 现有智能电网设备面临的安全形势非常严峻。

目前, 国内外已广泛开展智能设备网络威胁

监测分析工作。文献[7]基于属性的多源事件融合模型, 构建了源网荷系统恶意攻击关联分析方法; 文献[8]设计了基于自体集密度自动划分聚类方法, 开展自适应免疫计算的网络安全检测。同时, 也有采用度量方法对系统安全运行状态进行评估。文献[9]提供了一种基于网络流量评估、威胁评估、和漏洞评估 3 个模块的网络安全态势评估模型, 弥补了仅靠单一网络因素导致网络安全态势的评估不够全面的问题; 文献[10]通过建立网络信息系统的安全效用基线, 对网络系统的标识特征、流量特征、效用影响进行度量评估, 从而实现面向网络系统安全度量的攻击检测。上述方法在网络攻击威胁检测方面均有一定优势, 但在实际应用中存在以下问题: 与电力控制业务的结合深度不足, 导致检测方法效率不高; 无法对于操作系统运行过程中动态变化的内容进行度

收稿日期: 2020-08-09; 录用日期: 2020-12-18; 网络出版时间: 2021-01-21 15:40
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210121.1530.002.html
基金项目: 国网公司科技项目(5700-202024193A-0-0-00); 国家自然科学基金(U1736114)
* 通信作者. E-mail: zhuanzhuan2325@sina.com

引用格式: 吕卓, 郭志民, 陈岑, 等. 面向智能电网嵌入式设备的网络威胁动态评估方法[J]. 北京航空航天大学学报, 2022, 48(2): 324-330. LYU Z, GUO Z M, CHEN C, et al. A dynamic network threat evaluation method for smart grid embedded devices[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 324-330 (in Chinese).

量,检测准确率难以满足要求。

本文针对智能电网嵌入式设备安全防护方面存在的上述挑战,提出了一种基于组件动态可信用度的智能电网嵌入式设备网络攻击行为动态评估方法。相比于其他度量方法,该方法充分考虑了智能电网设备嵌入式系统“行为有限”和“状态有限”特点,在保证设备运行可靠性的前提下,通过对各种真实攻击行为的准确捕捉,快速、准确地对已知或未知的电力系统网络攻击进行实时动态的安全检测和评估,同时还能够实现网络攻击过滤装置的数据包标识和解析判别模式与文件配置模式的隔离,有效阻止攻击者通过网络接口模块对评估系统本身的攻击,提高设备自身安全性,实现对电网嵌入式系统的有效保护。

1 方法架构

1.1 方法整体架构

智能电网嵌入式设备网络攻击行为动态评估方法主要针对智能电网环境下各种嵌入式控制系统提供安全防护,该方法主要由网络接口模块、安全控制模块、嵌入式设备模拟机及安全分析服务器等部分组成。

运行时,由网络接口模块对当前通信网络行为的网络数据进行采集,利用安全控制模块对所接收的实际嵌入式设备发送的数据进行解析判别,在嵌入式设备模拟机中对网络行为的实际安全危害进行安全检测与评估,并通过安全分析服务器在平台配置属性、平台运行属性及用户认证属性 3 个方面进行多维度属性综合度量,最终得出安全检测结果。

智能电网嵌入式设备网络攻击行为动态评估方法的系统架构如图 1 所示。

本文在文献[11]所提出的智能电网嵌入式

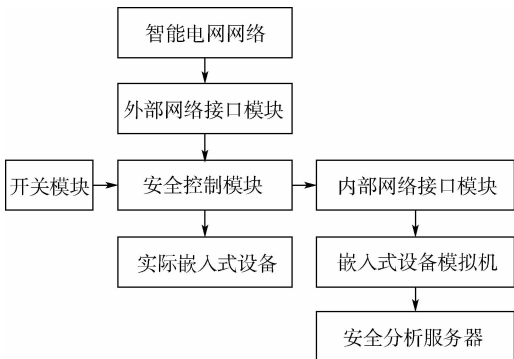


图 1 面向智能电网嵌入式设备的网络威胁动态评估方法架构

Fig. 1 Framework of dynamic network threat evaluation method for smart grid embedded devices

设备网络检测评估系统与检测评估方法的基础上,在安全控制模块中增加了解析判别模式与文件配置模式,通过 2 种模式物理切换实现了系统运行时内外部环境的隔离,有效阻止针对安全控制模块本身的攻击。以下对各模块具体功能进行说明。

1.2 方法模块设计

1.2.1 网络接口模块

网络接口模块用于完成 IP 数据包的收发,包括外部网络接口模块和内部网络接口模块。

1) 外部网络接口模块连接安全控制模块和智能电网网络,用于接收外部通信网络所发送的 IP 数据包并传送给安全控制模块,以及接收安全控制模块所发送的 IP 数据包并通过外部通信网络进行发送。

2) 内部网络接口模块分别连接安全控制模块、嵌入式设备模拟机、实际嵌入式设备,用于接收安全控制模块所发送的 IP 数据包并传送给实际嵌入式设备或嵌入式设备模拟机,以及接收实际嵌入式设备所发送的 IP 数据包并传送给安全控制模块。

1.2.2 安全控制模块

安全控制模块保存有与之相应目的地址和来源地址对应的密钥、发送序列号及接收序列号,用于对所接收的网络数据包进行解析判别。

安全控制模块还连接有开关模块,开关模块的信号输出端连接安全控制模块的信号输入端。开关模块用于向安全控制模块输入高电平或低电平信号,安全控制模块根据接收到开关模块发送的不同信号分别执行 IP 数据包标识和解析判别模式与文件配置模式。

文件配置模式下,安全控制模块仅通过串口通信模块与外界进行通信。当安全控制模块执行攻击过滤模式时,安全控制模块从内部启动,即从内部存储单元中读取 IP 数据包标识和解析判别程序并执行,安全控制模块不能访问外部存储器,能够保证不会篡改外部存储中的程序,进而保证了配置程序的安全。当安全控制模块执行配置程序运行模式时,安全控制模块从外部存储器中读取配置程序并执行,配置程序在用户的计算机中运行,用户计算机通过串口通信模块与安全控制模块进行通信,此时安全控制模块中运行配置程序,与用户的计算机形成 C/S 的工作模式。系统利用硬件开关隔离安全控制模块 2 种运行模式,安全控制模块执行配置程序时不通过网络进行,能够有效阻止通过网络接口模块发起的针对安全

控制模块本身的攻击,不论安全控制模块的配置程序有无漏洞,攻击者均不能修改安全控制模块的配置程序,安全性大为提高。

1.2.3 嵌入式设备模拟机

嵌入式设备模拟机用于对实际嵌入式设备的运行环境和计算环境进行模拟,包括硬件环境模拟和软件环境模拟,并对实际嵌入式设备受到网络攻击时其网络状态和主机状态变化进行检测,将网络状态及主机状态信息发送至安全分析服务器进行分析。

嵌入式设备模拟机中包括硬件可信平台模块(TPM),用于实现信息采集与组件动态可信度量^[12]。其中信息采集包括对于异常网络事件和主机事件的采集。异常网络事件信息包括异常的网络数据信息和网络流量信息,主机事件包括嵌入式设备模拟机配置信息和嵌入式设备模拟机运行信息^[13-15]。组件动态可信度量是利用嵌入式设备模拟机的特权控制机制分析嵌入式设备模拟机内的组件运行时的动态内存影像,从而对运行组件进行有效的动态度量,及时发现组件运行异常情况,对攻击做到主动防御,为嵌入式设备模拟机安全稳定运行提供安全可信的计算和运行环境^[16-18]。

1.2.4 安全分析服务器

安全分析服务器用于对嵌入式设备模拟机发送的网络状态及主机状态信息进行分析,通过基于平台配置属性度量、平台运行属性度量和用户认证属性度量进行的多维度属性综合度量,得出最终对外部攻击下的设备安全检测结果。

2 多维度属性综合度量方法

本文面向智能电网应用场景,基于文献[19-20]中提出的基于多维的环境信任度计算方法,在安全分析服务器中开展智能电网嵌入式设备的平台配置属性度量、平台运行属性度量、用户认证属性度量,最终综合以上3个方面度量结果,计算多维度属性综合度量结果,如图2所示。

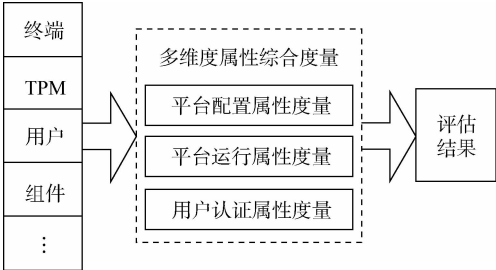


图2 多维度属性综合度量方法

Fig.2 Multi-dimensional attribute comprehensive measurement method

2.1 平台配置属性度量

平台配置属性度量反映了智能电网嵌入式设备各项配置的完整度,从而对遭受攻击时设备安全状态进行评估。本文通过将各组件完整性度量值存储在硬件可信平台模块TPM中,对其可信程度进行综合计算。

1) TPM采用可信密码技术,用于保证度量信息存储与传输过程的安全性。组件完整性度量值即存储在TPM中。

2) 各组件的完整性由安全分析服务器进行验证,安全分析服务器从TPM提供的安全通道中获取到完整性度量值,对其开展安全性验证。

3) 计算平台配置信任度 T_1 ,使用 $\{b_s, d_s, u_s\}$ 和 $\{b_f, d_f, u_f\}$ 分别表示完整性验证成功和失败的组件可信情况, b_s 表示该组件未受恶意代码影响的可能性, d_s 表示该组件受恶意代码影响的可能性, u_s 表示该组件受恶意代码影响的不确定程度; b_f 表示该组件对系统安全性造成破坏的可能性, d_f 表示该组件不会对系统安全性造成破坏的可能性, u_f 表示该组件对系统安全性是否造成破坏的不确定程度。使用 $\{b_i, d_i, u_i\}$ 表示平台配置信任度 T_1 :

$$\begin{cases} b_i = \kappa \left(\frac{1}{n} \sum_{i=1}^{n-f} b_{Si} \right) + \frac{1}{n} \sum_{j=1}^f b_{Fj} \\ d_i = \frac{1}{n} \sum_{i=1}^{n-f} d_{Si} + \frac{1}{n} \sum_{j=1}^f d_{Fj} \\ u_i = \frac{1}{n} \sum_{i=1}^{n-f} u_{Si} + \frac{1}{n} \sum_{j=1}^f u_{Fj} \end{cases} \quad (1)$$

式中: b_i 为完整性状态正常的概率; d_i 为完整性受到破坏的概率; u_i 为不确定度; n 为组件的个数; f 为完整性验证失败的组件数; κ 为调整因子,一般取 $(n-f)/n$,信任组件随着非信任组件的增多受影响越来越大。

2.2 平台运行属性度量

平台运行属性反映了嵌入式设备模拟机当前行为的信任属性。平台运行属性包括性能特性(如CPU、内存、硬盘使用情况和网络流量信息等)、可靠性(如成功率、丢包率和平均无故障时间等)和安全性(如非法连接次数、端口扫描次数和越权尝试企图等)。

平台运行属性度量通过将正常的网络通信事件作为正向事件,用 r 表示,将网络恶意攻击事件作为负向事件,用 z 表示。基于这些特性计算当前嵌入式设备模拟机运行情况的信任值 T_H ,其中 H 表示平台运行属性,平台运行属性信任度 T_H 由 $\{b_H, d_H, u_H\}$ 表示:

$$\begin{cases} b_H = \frac{r}{r+z+1} \\ d_H = \frac{z}{r+z+1} \\ u_H = \frac{1}{r+z+1} \end{cases} \quad (2)$$

式中: b_H 为正常网络通信的可能性; d_H 为非法网络通信事件的可能性; u_H 为正常网络通信的不确定程度。

2.3 用户认证属性度量

用户认证属性由其具有的身份凭证判定,并决定了其资源访问权限,是行为控制的基本属性。身份属性不同,在同样的平台上,可能造成安全风险也不同。当用户为了获得非法利益时,如访问未授权资源,可能利用系统漏洞或其他技术手段假冒其他用户身份,这就要求能对用户提交的身份凭证的可信性做出度量,即计算认证信任等级。在系统中,用户身份凭证可能有多种,如数字证书、指纹、虹膜乃至简单的 PIN 码,为了统一用户身份属性可信性的表达,用户认证属性度量采用认证方法被攻破的概率来计算认证信任级。用户认证属性度量问题实际上是计算多因素认证方式的破解概率问题。

用户认证属性度量具体步骤如下:

1) 假设认证方法 A 被攻击破解的概率是 $P(A)$, 则该认证方法 A 的可信等级 $\text{level}_A = -\lg(P(A))$ 。

2) 如果系统采取多因素认证方案, $A_1, A_2, \dots, A_m$, m 为认证因素的数量,如采用指纹、口令和证书 3 因素认证,那么 $m=3$; 则该多因素认证法被攻破的条件是全部认证方法均被攻破,其概率为 $P(A_1 \cap A_2 \cap \dots \cap A_m)$; 假设用户 U 通过了多因素认证,那么 U 通过系统认证后取得的可信等级 A_U 表示为

$$A_U = -\lg P(A_1 \cap A_2 \cap \dots \cap A_m) \quad (3)$$

2.4 多维度属性综合度量

基于平台配置属性度量、平台运行属性度量和用户认证属性度量进行多维度属性综合度量是在用户认证属性度量的基础上,对平台配置属性度量和平台运行属性度量的加权平均。设 α_1 和 α_H 分别为平台配置属性度量和平台运行属性度量的权重, $\alpha_1 + \alpha_H = 1$, 则嵌入式设备模拟机安全度量评估值 $T_p = \{b_p, d_p, u_p\}$ 为

$$\begin{cases} b_p = \alpha_1 b_1 + \alpha_H b_H \\ d_p = \alpha_1 d_1 + \alpha_H d_H \\ u_p = \alpha_1 u_1 + \alpha_H u_H \end{cases} \quad (4)$$

式中: b_p 为实际嵌入式设备模拟机安全可信的可能性; d_p 为实际嵌入式设备模拟机非安全可信的可能性; u_p 为实际嵌入式设备模拟机安全可信的不确定程度。

基于式(4),即可得出最终安全评估结果。

3 实验结果分析

3.1 实验环境搭建

利用配电自动化模拟主站及真实 DTU、FTU 配电自动化终端搭建了一个配电自动化系统模拟仿真环境,并对攻击行为检测模型进行了系统实现,完成检测系统与配电自动化系统的防护接入,如图 3 所示。

通过本文设计的多维度属性综合度量方法,对工具扫描、数据篡改、报文注入等一系列模拟攻击行为进行评估,通过平台配置、平台运行、用户认证等维度的度量结果,对攻击行为进行实时检测,并通过攻击检测准确率和业务响应时间来检验该模型对于攻击的实际检测效果。

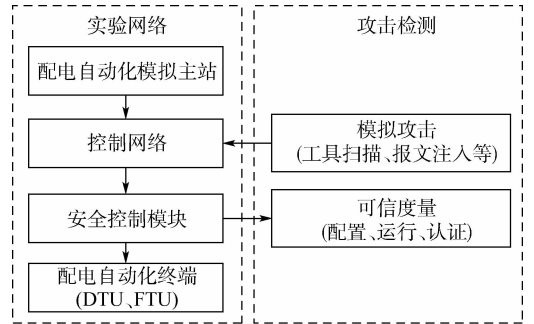


图 3 实验网络结构

Fig.3 Experimental network structure

3.2 实验过程分析

1) 攻击模拟。将自动化渗透工具接入配电自动化实验网络,工具中预置端口扫描、畸形报文、消息篡改、拒绝服务等多类安全攻击测试脚本,对配电自动化终端设备开展不同类型的攻击测试。

2) 可信度量。基于嵌入式设备虚拟机与安全分析服务器,依托本文设计度量方法,对网络攻击行为及该攻击行为对嵌入式设备虚拟机内配置状态、运行状态、用户状态造成的影响进行度量,确定不同攻击模型下的度量评估影响,并根据综合度量结果对于攻击行为进行判断评估。

3) 结果呈现。本文设计的评估方法的目标是检测智能电网嵌入式设备中的攻击行为,因此本文对不同攻击类型及不同攻击频率下的评估结果进行分析,并以攻击行为检测准确率与检测响

应时间作为衡量本文所提评估方法的重要指标。同时将本方法实验结果与当前主流的入侵检测算法 Apriori 进行对比分析,以验证本文方法的有效性和准确性。

其中,Apriori 算法是一种挖掘关联规则的频繁项集算法,该算法被广泛应用在网络入侵检测中,具有较好的效果。但随着网络攻击逐渐向定值化方向发展,传统 Apriori 算法只对网络层事件进行关联,存在关联精度不高的不足。

3.3 实验数据对比

评估效果通过攻击检测准确率和响应时间 2 个方面来体现。如图 4 所示,由于配电自动化系统通信过程的行为有限和状态有限特点,攻击行为给嵌入式设备所带来的状态变化是非常明显的,对于真实攻击行为的检测准确率比较高,同时由于本文方法结合了智能电网设备业务特点,其检测准确率、响应时间与 Apriori 算法相比均具有一定的优势。

如图 5 所示,当减缓攻击动作的发生频率时,检测准确率依然保持在较高的水平,但是由于单

位时间窗内生成的攻击行为样本缩小,响应时间出现了一定的下滑,但是仍然能够给出较为及时的安全响应,在此种条件下,Apriori 算法由于其依托的安全事件关联性下降,其检测指标同样产生下滑,本文设计方法依然保持一定的检测优势。

与此类算法相比,本文方法将攻击行为与电网真实业务场景相结合,将攻击影响与智能设备真实运行环境相结合,针对典型控制系统攻击类型进行检测对比分析,相比当前常用入侵检测方法,显著提升了针对智能电网设备的检测准确率。

基于以上实验环境,针对本文所设计的安全控制模块自身安全性开展攻击测试,以验证方法自身抵抗外部攻击的防护能力。

如表 1 所示,利用命令执行、畸形数据、内存溢出、内核提权及拒绝服务等攻击手段对于本文设计安全控制模块进行攻击,由于本文方法采用了内外部存储隔离、网络接口隔离等安全控制方法,能够对外部网络进行有效的边界隔离,使得攻击行为无法访问到真实控制系统,即使在核心模块存在漏洞的情况下,不会引入新的风险点,保证动态度量系统自身的安全性,满足智能电网嵌入式设备自身的安全防护需求。

表 1 攻击防御测试

Table 1 Attack protection test		
攻击类型	攻击样本数量	成功次数
命令执行	1 103	0
畸形数据	476	0
内存溢出	234	0
内核提权	16	0
拒绝服务	18 920	0

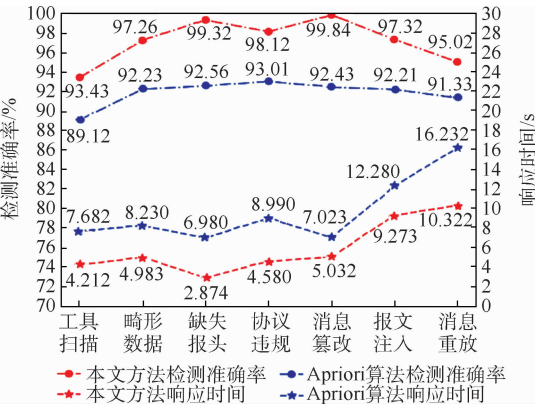


图 4 攻击行为检测模型检测效果(高攻击频率)

Fig. 4 Detection result of attack behavior detection model (high attack frequency)

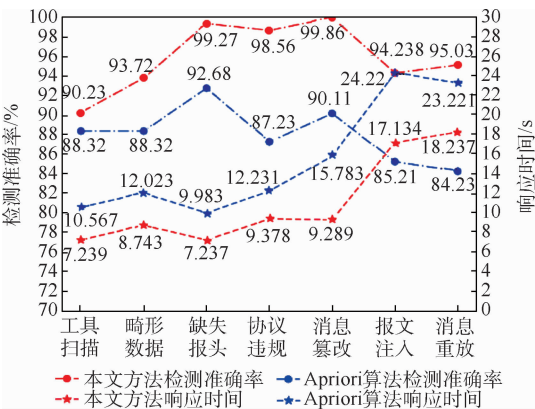


图 5 攻击行为检测模型检测效果(低攻击频率)

Fig. 5 Detection result of attack behavior detection model (low attack frequency)

4 结 论

本文提出了一种智能电网嵌入式设备网络攻击行为动态评估方法和系统,能够实现对嵌入式设备网络攻击行为的实时动态评估,同时能够更加真实反映攻击行为对设备状态与业务运行带来的影响,主要具备以下特点:

- 1) 在应对常规网络攻击方面,本文方法能够保持很高的检测准确率与较低的响应时间。
- 2) 在面临针对电网业务的隐蔽性攻击行为方面,由于攻击者普遍会降低攻击强度和攻击频次,普通检测方法难以保持较好检测效果。本文方法充分结合电网业务场景,检测准确率和响应时间始终保持在较高水平。
- 3) 本文系统在提供高效检测能力的同时,通过引入内外部存储隔离、网络接口隔离等安全控

制机制,充分保证了自身系统的安全性,避免了引入新的安全风险。

智能电网设备承载业务众多,未来将针对不同电网业务对方法进一步优化,提升针对智能电网的全面防护能力。

参考文献 (References)

[1] 何金栋,王宇,赵志超,等. 智能变电站嵌入式终端的网络攻击类型研究及验证[J]. 中国电力,2020,53(1): 81-91.
HE J D,WANG Y,ZHAO Z C,et al. Type and verification of network attacks on embedded terminals of intelligent substation [J]. Electric Power,2020,53(1):81-91 (in Chinese).

[2] 梅文明,李美成,孙炜,等. 一种面向分布式新能源网络的终端安全接入技术[J]. 电网技术,2020,44(3):953-961.
MEI W M,LI M C,SUN W,et al. Terminal security access technology for distributed new energy networks[J]. Power System Technology,2020,44(3):953-961 (in Chinese).

[3] 王宇,李俊娥,周亮,等. 针对嵌入式终端安全威胁的电力工控系统自愈体系[J]. 电网技术,2020,44(9):3582-3594.
WANG Y,LI J E,ZHOU L,et al. A self-healing architecture for power industrial control systems against security threats to embedded terminals[J]. Power System Technology,2020,44(9):3582-3594 (in Chinese).

[4] 李田,苏盛,杨洪明,等. 电力信息物理系统的攻击行为与安全防护[J]. 电力系统自动化,2017,41(22):162-167.
LI T,SU S,YANG H M,et al. Attacks and cyber security defense in cyber-physical power system[J]. Automation of Electric Power Systems,2017,41(22):162-167 (in Chinese).

[5] 周敦凯. 基于可信计算的嵌入式系统安全设计[J]. 自动化与仪器仪表,2020(10):81-84.
ZHOU D K. Security design of embedded system based on trusted computing[J]. Automation & Instrumentation,2020(10):81-84 (in Chinese).

[6] FRAGOULI C,PRABHAKARAN V M,CZAP L,et al. Wireless network security: Building on erasures[J]. Proceedings of the IEEE,2015,103(10):1826-1840.

[7] 章锐,费稼轩,石聪聪,等. 特定攻击场景下源网荷系统恶意攻击关联分析方法[J]. 中国电力,2019,52(10):1-10.
ZHANG R,FEI J X,SHI C C,et al. Malicious attack correlation analysis method of source-grid-load system under specific attack scenarios[J]. Electric Power,2019,52(10):1-10 (in Chinese).

[8] 陈晋音,徐轩衍,苏蒙蒙. 基于自适应免疫计算的网路攻击检测研究[J]. 计算机科学,2018,45(S1):364-370.
CHEN J Y,XU X Y,SU M M. Research on network attack detection based on self-adaptive immune computing[J]. Computer Science,2018,45(S1):364-370 (in Chinese).

[9] 李小雨,王怀彬. 基于 Renyi 交叉熵与 CVSS 的网络安全态势评估模型[J]. 天津理工大学学报,2019,35(5):12-17.
LI X Y,WANG H B. Thenetwork security situation assessment

model based on Renyi cross entropy and CVSS[J]. Journal of Tianjin University of Technology,2019,35(5):12-17 (in Chinese).

[10] 曾颖明,谢小权,吴明杰,等. 一种面向网络系统安全度量的攻击检测方法:CN109167794B[P]. 2021-05-14.
ZENG Y M,XIE X Q,WU M J,et al. An attack detection method for network system security measurement:CN109167794B[P]. 2021-05-14 (in Chinese).

[11] 莫坚松,张之刚,牛霜霞,等. 智能电网嵌入式设备网络检测评估系统与检测评估方法:CN103905450B[P]. 2017-05-31.
MO J S,ZHANG Z G,NIU S X,et al. Smart grid embedded equipment network detection and evaluation system and detection and evaluation method:CN103905450B[P]. 2017-05-31 (in Chinese).

[12] WANG W,LIU J Q,PITSILIS G,et al. Abstracting massive data for lightweight intrusion detection in computer networks[J]. Information Sciences,2018,433-434:417-430.

[13] WANG W,SHANG Y Y,HE Y Z,et al. BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors[J]. Information Sciences,2020,511:284-296.

[14] LI Y D,ZHANG L,LV Z,et al. Detecting anomalies in intelligent vehicle charging and station power supply systems with multi-head attention models[J]. IEEE Transactions on Intelligent Transportation Systems,2021,22(1):555-564.

[15] WANG W,WANG X,FENG D W,et al. Exploring permission-induced risk in Android applications for malicious application detection[J]. IEEE Transactions on Information Forensics and Security,2014,9(11):1869-1882.

[16] WANG W,LI Y Y,WANG X,et al. Detecting Android malicious apps and categorizing benign apps with ensemble of classifiers[J]. Future Generation Computer Systems,2018,78:987-994.

[17] WANG X,WANG W,HE Y Z,et al. Characterizing Android apps' behavior for effective detection of malapps at large scale[J]. Future Generation Computer Systems,2017,75:30-45.

[18] WANG W,ZHAO M C,GAO Z Z,et al. Constructing features for detecting android malicious applications: Issues, taxonomy and directions[J]. IEEE Access,2019,7:67602-67631.

[19] 常朝稳,徐江科. 终端行为可信评估及其访问控制方法研究[J]. 小型微型计算机系统,2014,35(3):493-499.
CHANG C W,XU J K. Research on behavior assessment and access control for terminal[J]. Journal of Chinese Computer Systems,2014,35(3):493-499 (in Chinese).

[20] 徐江科. 信息系统终端用户行为可信评估研究[D]. 郑州: 解放军信息工程大学,2012.
XU J K. Research on trusted assessment of user behavior on information system terminal[D]. Zhengzhou: PLA Information Engineering University,2012 (in Chinese).

A dynamic network threat evaluation method for smart grid embedded devices

LYU Zhuo^{1,*}, GUO Zhimin¹, CHEN Cen¹, MO Jiansong¹, CHANG Chaowen²

(1. State Grid Henan Electric Power Research Institute, Zhengzhou 450052, China;

2. Information Engineering University of PLA Strategic Support Force, Zhengzhou 450004, China)

Abstract: Due to the limited computing and storage resources, the smart grid embedded devices cannot deal with the network attacks effectively and the security assessment method is weak. In order to solve these problems, a dynamic network attack behavior evaluation method for smart grid embedded devices is proposed. This method uses the security control module to analyze the communication data stream of the actual embedded device, and conduct security detection evaluation of the impact of the attack behavior in the embedded system simulator by using the component dynamic trust measurement. The final security evaluation result of the network attacks is obtained based on the whole process dynamic comprehensive measurement of the platform configuration property, the platform operation attribute and the user authentication attribute. The method is tested in the actual environment of the power distribution automation system and the power utilization information collection system. The results show that, aimed at the common attacks against the embedded devices, accuracy rate of the proposed detection method can reach more than 90%. This method provides good safety assessment accuracy, and meanwhile achieves effective upgrade of its own security.

Keywords: dynamic evaluation; attribute measurement; security detection; network attack behavior; dynamic trust measurement

Received: 2020-08-09; **Accepted:** 2020-12-18; **Published online:** 2021-01-21 15:40
URL: kns.cnki.net/kcms/detail/11.2625.V.20210121.1530.002.html
Foundation items: State Grid Technology Project (5700-202024193A-0-0-00); National Natural Science Foundation of China (U1736114)
*** Corresponding author.** E-mail: zhuanzhuan2325@sina.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0397

基于 LIME 的恶意代码对抗样本生成技术

黄天波, 李成扬, 刘永志, 李煜辉, 文伟平*

(北京大学 软件与微电子学院, 北京 102600)

摘 要: 基于机器学习检测恶意代码技术的研究和分析, 针对机器学习模型对抗样本的生成提出一种基于模型无关的局部可解释(LIME)的黑盒对抗样本生成方法。该方法可以对任意黑盒的恶意代码分类器生成对抗样本, 绕过机器学习模型检测。使用简单模型模拟目标分类器的局部表现, 获取特征权重; 通过扰动算法生成扰动, 根据生成的扰动对原恶意代码进行修改后生成对抗样本; 基于 2015 年微软公布的常见恶意样本数据集和收集的来自 50 多个供应商的良性样本数据对所提方法进行实验, 参照常见恶意代码分类器实现了 18 个基于不同算法或特征的目标分类器, 使用所提方法对目标分类器进行攻击, 使分类器的真阳性率均降低到接近 0。此外, 对 MalGAN 和 ZOO 两个先进的黑盒对抗样本生成方法与所提方法进行对比, 实验结果表明: 所提方法能够有效生成对抗样本, 且方法本身具有适用范围广泛、能灵活控制扰动和健全性的优点。

关 键 词: 对抗样本; 恶意代码; 机器学习; 模型无关的局部可解释(LIME); 目标分类器

中图分类号: TP309

文献标志码: A 文章编号: 1001-5965(2022)02-0331-08

近年来, 恶意代码检测技术的研究表明机器学习在代码检测问题上被越来越多的研究人员应用, 众多学者提出将机器学习技术作为下一代恶意代码分类器的关键组成部分^[1-5]。分类器从恶意代码中提取特征, 使用机器学习算法对良性程序与恶意程序进行分类。根据使用的特征性质, 可以将恶意代码检测技术分为静态检测^[6-7]和动态检测^[8]。然而机器学习分类器在多个领域都被证明是不安全的。随着机器学习越来越广泛的应用, 对抗样本攻击和防御的研究也在变得越来越有意义, 这一领域通常被归为对抗机器学习(adversarial learning)。在恶意代码检测问题上, 基于机器学习的分类器在面对对抗样本攻击时也

可能是非常脆弱的。因此, 研究在恶意代码检测问题上的对抗样本生成技术可以增加对这种攻击方法的了解, 避免将分类器暴露在这种攻击之下, 从而针对该攻击提出针对性或者普适性的防御方法。同时, 研究人员可以通过攻击方法攻击其所研究的分类器, 对分类器的鲁棒性进行评估。

本文提出了一种基于模型无关的局部可解释(local interpretable model-agnostic explanations, LIME)^[9]的对抗样本生成方法。该方法可针对未知算法和参数细节的分类器生成有效的对抗样本, 并且通过引入扰动常量, 使该方法具有了适用范围广泛、灵活控制扰动大小的优点。同时, 在实验中验证了该方法的健全性。

收稿日期: 2020-08-09; 录用日期: 2020-10-23; 网络出版时间: 2021-01-21 09:57
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210121.0934.001.html
基金项目: 国家自然科学基金(61872011)
* 通信作者. E-mail: weipingwen@ss.pku.edu.cn

引用格式: 黄天波, 李成扬, 刘永志, 等. 基于 LIME 的恶意代码对抗样本生成技术[J]. 北京航空航天大学学报, 2022, 48(2): 331-338. HUANG T B, LI C Y, LIU Y Z, et al. Adversarial sample generation technology of malicious code based on LIME[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 331-338 (in Chinese).

1 相关研究

为使提出的对抗样本生成方法不仅可攻击分类器,同时有助于评估防御^[10]、提供具体级别的安全产品^[11],针对对抗样本相关技术、攻击者对抗能力的描述和样本修改进行了相关的研究。

1.1 恶意代码对抗样本相关技术

对抗样本攻击意为通过微小地修改机器学习分类器的输入样本,诱导分类器对修改后的输入产生错误的输出结果。2014 年,Szegedy 等^[12]在神经网络模型背景下给出了严谨的问题描述,之后出现了基于梯度的白盒解决方案^[13]。但是因为现实中的大部分分类器都是闭源的,相较于白盒方法,黑盒更具有广泛的使用价值。

2017 年,Hu 和 Tan^[14]提出了 MalGAN,其是一种针对恶意代码分类器的黑盒对抗样本生成方法。使用该方法可以对一些使用 One-Hot 型(只有 0 和 1 两种取值)特征的黑盒分类器生成对抗样本。

2017 年,Chen 等^[15]提出零阶优化(zeroth-order optimization,ZOO)方法,其是一种基于零阶优化估计目标分类器梯度进而生成对抗样本的方法。由于不需要梯度,ZOO 也是一种黑盒方法,使用 ZOO 无需额外训练模型。

2016 年,Ribeiro 等^[9]提出 LIME 方法,其是一种与模型无关的方法(model-agnostic)。原理是:使用可解释的简单模型在局部逼近目标模型,无论目标模型有多复杂,其在局部(可理解为切点)的趋势(可理解为切线)都可以用一个简单模型来刻画,从而通过简单模型解释局部的特征权重。对于给定的目标模型 f 和输入向量 $\mathbf{x}$,LIME 方法可以在 f 的局部使用简单模型模拟目标模型的局部性质,从而判断 $\mathbf{x}$ 各分量对分类结果的影响权重。局部模拟需要一组与 $\mathbf{x}$ 相近的特征,使用 $\mathbf{z}$ 代表临近特征。LIME 对特征的解释 $\xi(\mathbf{x})$ 通过下式表达:

$$\xi(\mathbf{x}) = \arg \min_{g \in G} L(f, g, \Pi_{\mathbf{x}}) + \Omega(g) \tag{1}$$

式中: $g \in G$, G 为用于模拟目标模型的简单模型集合,通常包括线性模型和决策树; $\Pi_{\mathbf{x}}(\mathbf{z})$ 用于度量局部特征 $\mathbf{z}$ 和 $\mathbf{x}$ 的接近程度; $\Omega(g)$ 用来衡量 g 的复杂性(不利于解释的程度),通常取决于选用的简单模型,如选用决策树,决策树的深度就是影响 $\Omega(g)$ 的主要因素; L 为损失函数,用来描述 g 在 $\mathbf{x}$ 附近模拟 f 的效果,模拟效果越好,则 L 的值越小, L 的一种实现如下:

$$L(f, g, \Pi_{\mathbf{x}}) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathbf{Z}} \Pi_{\mathbf{x}}(\mathbf{z}) (f(\mathbf{z}) - g(\mathbf{z}'))^2 \tag{2}$$

在 L 中, $\mathbf{z}$ 和 $\mathbf{x}$ 的距离越小,则 $\Pi_{\mathbf{x}}(\mathbf{z})$ 越大,也就是说,离 $\mathbf{x}$ 越近,误差权重越大,通过最小化,这样的损失函数使 g 在 $\mathbf{x}$ 局部获得较好的模拟效果,而 $L(f, g, \Pi_{\mathbf{x}}) + \Omega(g)$ 最小,意味着使局部模拟和可解释的综合效果最好。Ribeiro 等^[9]用 K-Lasso 算法选择 k 个特征(通过正则化路径),基于对 $\mathbf{x}$ 的随机扰动生成一组数据,通过最小化 $L(f, g, \Pi_{\mathbf{x}}) + \Omega(g)$ 学习到最优的简单模型 g ,通过对 g 的参数 $\mathbf{w}$ 分析即可得到特征权重,如当 g 是线性模型时, g 的参数 $\mathbf{w}$ 就对应了每个特征的权重。如果 k 被设置为特征的总数,那么将得到所有特征的权重。由于 LIME 方法在设计时没有对目标模型做任何假设约束,使用时无需知道目标模型的算法和参数,理论上可用于解释任意黑盒模型。因为只在局部用简单模型模拟,效果好速度快,所以这也是本文所设计的对抗样本方法选用 LIME 的原因。

1.2 对攻击者能力的描述

精确的描述攻防场景,需要对攻击者的攻击能力做出合理的预估^[16]。本文采用 Stokes 等^[11]的方法,将特征分为正特征(positive feature)和负特征(negative feature),正特征表示样本中有利于使分类器判断为恶意代码的特征,这些特征往往代表恶意行为,负特征表示样本中有利于使分类器判断为良性代码的特征。Crandall 等^[17]指出,攻击者通常采取变构策略,使用替代代码的方式来达到所需的恶意目标,攻击者有能力删除正特征或添加负特征。在 MalGAN 方法中,假设攻击者只能添加特征,不能删除特征,从而保证不影响原样本的程序特征,但该方法不对添加特征的数量进行限制,本身的假设存在低估攻击者对抗能力的可能。Incer 等^[18]使用了特定的对抗能力来描述其所提供的安全机制的安全边界,通过分析一系列特征的修改,并评估每种类型的修改是否简单廉价,从而确定攻击者的对抗能力。但是在对这些类别的修改进行评估时,评估的标准缺少灵活性。事实上,对于攻击能力较强的专业黑客,即使是被认为困难的修改,也是可以完成一定数量的,而对于被认为是简单廉价的修改,在现实场景中,攻击者也不一定会不计数量的大量修改,大量未使用的特征可能增大被检测为对抗样本的概率,如 Xu 等^[19]提出一种检测对抗样本的方法,筛选被添加到原样本中却未使用到的修改作为检测标准之一。

基于上述分析,可以明显看到这些假设容易低估或者高估一些攻击者,导致只能用于特定研究背景,具有一定的局限性。为了寻找适合衡量对抗能力的方法,需要进一步分析攻击者修改样本的过程。

1.3 样本修改过程

攻击者在其能力和成本范围内对原样本进行修改,在保证新样本可执行的前提下,要确保新样本与原样本主要程序功能相同,即保证对抗样本的有效性。一般修改的过程分为 2 个部分:逆提取和逆预处理。逆提取中,针对新旧样本的差异特征 r 和中间层表示的对应关系,通过直接修改中间层文件或者样本的源代码以满足在不影响其他特征的基础上添加新特征。逆预处理实现修改

后的中间层文件到可执行文件的转变,可以通过逆向工程的方式从中间文件转换为可执行文件,甚至若在逆提取阶段中直接修改源码,则可编译得到可执行文件。

通过对攻击过程进行分析可知,修改样本的难度和成本主要体现在逆提取过程中。因此引入扰动常量的概念,用于描述攻击者的对抗能力,并提出基于 LIME 的扰动方式的实现。

2 对抗样本生成方法设计

2.1 基于 LIME 的恶意代码对抗样本生成过程

对抗样本的生成过程主要包含 4 个部分:探测(A 部分)、扰动算法(B 部分)、逆向过程(C 部分)和验证部分(最下方虚线部分),如图 1 所示。

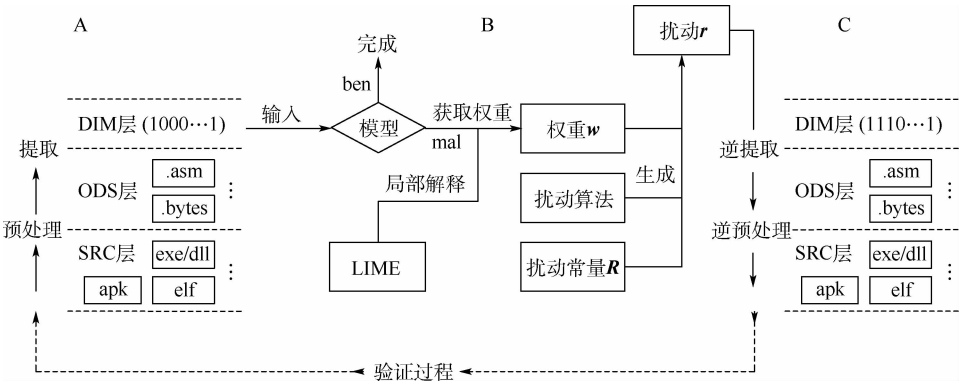


图 1 基于 LIME 的恶意代码分类器对抗样本生成过程

Fig. 1 Adversarial sample generation process of malicious code classifier based on LIME

本文对抗样本生成方法所解决的问题可描述为

$$\begin{cases} f(\boldsymbol{x}) = \text{mal} \\ f(\boldsymbol{x} + \boldsymbol{r}) = \text{ben} \end{cases} \quad (3)$$

$$g(\boldsymbol{x}) = g(\boldsymbol{x} + \boldsymbol{r}) \quad (4)$$

式中:mal (malicious) 为恶意代码分类标签;ben (benign)为良性代码分类标签; $\boldsymbol{x}$ 为目标分类器 f 的输入,表示恶意代码中提取的特征; $\boldsymbol{r}$ 为对 $\boldsymbol{x}$ 的扰动 (perturbation);函数 g 为获取样本的主要程序功能, $g(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 原本的主要程序功能, $g(\boldsymbol{x} + \boldsymbol{r})$ 为修改后的主要程序功能。在保证代码语义的前提下,诱使目标分类器将恶意代码标识为良性代码。对图 1 的 4 个模块做出具体阐述如下:

- 1) 探测。通过特征工程获取样本特征,探测目标分类器 f 的分类结果,当结果为 ben 时结束方法,否则进行下一步。
- 2) 扰动。使用基于 LIME 的扰动方法生成一个扰动 $\boldsymbol{r}$,满足 $f(\boldsymbol{x} + \boldsymbol{r}) = \text{ben}$,生成成功则进入下一步,失败则方法以失败结束。
- 3) 逆向过程。根据扰动 $\boldsymbol{r}$ 修改相应样本程

序,生成最终可逃逸检测的对抗样本,使用该对抗样本再次进行步骤 1 的探测,应能够正常结束方法,否则可能是在逆向过程出错,检查错误后重试方法。

- 4) 验证过程。如果修改后的样本满足式(4),且能够按照探测模块的方法再次探测,并能得到探测结果 ben,则验证通过,结束方法。

2.2 扰动算法设计

扰动模块的实现包括 2 个方面:扰动常量和扰动方式。扰动常量形式化地描述了攻击者在有限成本内修改样本,且保持样本主要功能不变的能力。扰动常量越大,攻击者能力越强;扰动方式指定了对于具体特征的修改,不同于使用雅可比算法^[20]等白盒方法,使用 LIME 模拟黑盒分类器在恶意样本处的局部表现,从而确定影响样本分类的关键特征,然后使用一个扰动算法,在保证不影响输入样本主要程序功能的同时,修改部分关键特征。

2.2.1 扰动常量

扰动常量 $\boldsymbol{R}$,用于描述在特定攻击场景下攻

击者的成本和能力边界,也就是对抗能力——攻击者在有限成本内修改样本,且使修改后的样本保持原样本主要程序性质不变的能力。以下是 R 的一般形式:

$$R = \{(d_i, s_{i1}, s_{i2}) \mid s_{i1} \leq s_{i2}, d_i \in [0, 1]\}_{i=1}^k \quad (5)$$

式中: k 为正整数; d_i, s_{i1} 和 s_{i2} 均为实数。 R 为一个 $k \times 3$ 的向量,包含 k 个扰动规则; i -th 表示第 i 个扰动规则,通常用 1 个或 2 个扰动规则描述一种特征,对应添加和删除规则。不妨假设 i 规则描述一种 A 特征,则攻击者使用 i 规则对 A 特征所能修改的维数比例的最大值用 d_i 表示, d_i 取值在 $[0, 1]$ 之间,如果 A 特征在特征向量中共 m 维,攻击者真实修改的 A 特征维数应不大于 $m \times d_i$;用 s 描述对具体某维特征的修改能力,允许对某个特征在初始值的基础上加 s ,这里的 s_{i1} 和 s_{i2} 用来描述 s 的取值范围。当 s_{i1} 与 s_{i2} 确定时,区间 $[s_{i1}, s_{i2}]$ 表示对所有 A 特征, s 取值应在 $[s_{i1}, s_{i2}]$ 范围内。这个区间往往由特征的性质、攻击者的攻击成本和能力所决定。这样 k 个扰动规则综合起来,就描述了 R 表示的范围。攻击者能力越强或越不计成本, R 表示的范围就越大;特征越容易修改, R 表示的范围也越大。

2.2.2 基于 LIME 的扰动方式实现

扰动方式指定了对于具体特征的修改。首先使用 LIME 方法求解出逼近真实情况的 ω 。在 LIME 方法中,需要选择简单模型 g 来模拟目标分类器,这里以线性模型为例,那么 $g(x)$ 可以表示为

$$g(x) = \omega_1 \cdot x_1 + \omega_2 \cdot x_2 + \omega_3 \cdot x_3 + \cdots + \omega_m \cdot x_m \quad (6)$$

式中: $\omega = \{\omega_1, \omega_2, \omega_3, \cdots, \omega_m\}$ 为 g 的参数。使用 K-Lasso 算法选择 k 个特征(通过正则化路径),然后基于对 x 的随机扰动生成一组临近 x 的数据,在这组数据基础上最小化损失函数 $L(f, g, \Pi_x)$,可以得到此时 g 的参数 ω ,用 ω 近似表示特征 x 的权重,即令 $w = \omega$,从而求解出 w ,如果 k 被设置为特征的总数,那么将得到所有特征的权重。

当攻击者获取权重 w 之后,需要根据 w 和扰动算法生成一个符合式(3)、式(4)的扰动 r ,且攻击者可以在对抗能力范围内对恶意样本做出 r 对应的修改。不妨设分类器给出的置信度在 $[0, 1]$ 区间内,以 0.5 为分界,0 表示百分之百确定是 ben 样本,1 表示百分之百确定是 mal 样本。创建扰动 r 的过程如算法 1 所示。

算法 1 创建扰动 r 的算法。

Data: $w = \{w_1, w_2, w_3, \cdots, w_m\}$, distance, R

```
1   $r \leftarrow \{\Delta x_1, \Delta x_2, \Delta x_3, \cdots, \Delta x_m\}, \Delta x_i = 0$ 
2   $U \leftarrow \{(1, w_1), (2, w_2), (3, w_3), \cdots, (m, w_m)\}$ 
3  sum  $\leftarrow 0$ 
4  list  $\leftarrow \text{sort}(|w|)(U)$  // 根据  $w_i$  的绝对值进行递增排序
5  if not list then
6      return fail // list 为空
7  end
8  while list do
9       $i, w \leftarrow \text{list.remove}(0)$  // 移除第 1 个元素
10      $\Delta x_i \leftarrow \arg \min_{\Delta x, \Delta x \in R} \Delta x \cdot w$ 
11     sum  $\leftarrow \text{sum} + \Delta x_i \cdot w$ 
12     if sum < distance then
13         break
14     end
15 end
16  $f(x + r) \leftarrow \text{ben? return } r$ ; distance  $\leftarrow \text{distance} \times 2$ , go 5
```

攻击者首先探测到分类器 f 检测恶意样本 x 的置信度 $\text{confidence} = f(x)$,对于恶意样本, confidence 在 $(0.5, 1]$ 之间。由于 w 是已经求得的,攻击者需要在 x 中选择一些维度,在使 confidence 减少的方向上,做一些能力和成本范围之内的改变。考虑 g 取线性模型的情况,此时有 $g(x) = w \cdot x$,且 $g(x + r) = g(x) + g(r)$,因为 $g(x)$ 是对 $f(x)$ 的局部模拟,攻击者可以通过 g 近似计算 confidence ,所以一般满足 $g(r) = w \cdot r < -0.2$ 即可。在具体选择 r 的维度时,由于 w 已经用 LIME 方法求得,可以使用贪心法,按权重由大到小,依次判断能否对相应特征做出符合 R 要求的修改,从而挑选合适的维度,用 $\Delta x \in R$ 表示对某个特征进行 Δx 的变动是符合 R 要求的修改,如果符合,就加入候选维度。

3 实验评估

为了验证本文所提出的基于 LIME 的对抗样本生成方法的效果,分别进行攻击实验和对比实验。攻击实验用于测试本文方法在目标分类器上的攻击效果;对比实验用于和同类方法做比较,增强结论的说服力。本节将介绍实验环境、数据集、评估指标、目标分类器设置以及实验设计。

3.1 实验环境

实验的软硬件环境信息如表 1 所示,实验代码主要使用 Python3.6.3 实现,在数据集准备环节

表 1 实验的硬件、软件环境	
Table 1 Hardware and software environment of experiment	
软/硬件环境	具体信息
硬件环境	内存:16 GB
	CPU:Inter(R) Core(TM) i7-8550U
软件环境	IDA pro 7.0
	lime 0.1.1.37
	keras 2.3.1
	Python 3.6.3
	tensorflow 1.15.0
	numpy 1.18.1
	sklearn 0.20.0
	adversarial-robustness-toolbox 1.1.1

借助了 IDA pro 7.0 提供的 Python 2.7.17 脚本执行接口,在 IBM 提供的 adversarial-robustness-toolbox 1.1.1包中,封装了许多应用于机器学习分类器的攻击与防御方法,这包括本文要使用的 ZOO 方法,此外本文所要使用的 LIME 方法被封装在 lime 0.1.1.37 包中。

3.2 数据集

本文收集了 2 组不相交的 Win32 PE 文件,这 2 组文件的收集方式借鉴于文献[21],对于每个 PE 文件,都使用其 IDA pro 反汇编生成的 ASM 文件来代表 PE 文件。良性样本源于系统镜像中使用 ninite^[22]安装的 50 多个供应商的应用软件,进而有效避免学习和识别与特定供应商相关联的文件^[23]。恶意样本包含来自 Ramnit、Lollipop 等 9 个恶意代码家族的 10 868 个 Win32 PE 恶意程序对应的 ASM 文件,这些文件来自 2015 年微软举办 Kaggle 比赛^[24]时公开的数据。

3.3 评估指标

因为攻击者希望原本被识别的恶意样本,经改变为对抗样本后,被标识成良性样本,攻击前后的真阳性率 TPR 之差就是有效对抗样本的比例。为了直观考虑,将攻击前后的 TPR 之差与攻击前的 TPR 之比称为攻击成功率,表示为 ASR,则有 $ASR = 1 - TPR_{after} / TPR_{before}$;除上述 2 个指标外,为了保证说服力,本文采用准确率 ACC 来评估自建的目标分离器,仅保留那些准确率 90% 以上的分类器纳入实验。采用的评估指标如表 2 所示。表中:TP 为真阳性样本数量,FP 为假阳性样本数量,FN 为假阴性样本数量,TN 为真阴性样本数量。

为充分测试方法的效果,目标分类器根据使用的算法或特征差异,可分为 18 个,如表 3 所示。算法涵盖线性、树形和深层神经网络类算法,包括 LR、RF、SVM、MLP 算法,特征包括 API、opc-2gram、opc-3gram。

表 2 评估指标	
Table 2 Evaluation indicators	
评估指标	公式
真阳性率 TPR	$TPR = TP / (TP + FN)$
准确率 ACC	$ACC = (TP + TN) / (TP + TN + FP + FN)$
攻击成功率 ASR	$ASR = 1 - TPR_{before} / TPR_{after}$

表 3 目标分类器设置			
Table 3 Target classifier setting			
算法	API	opc-2gram	opc-3gram
LR	#1	#2	#3
RF	#4	#5	#6
SVM	#7	#8	#9
MLP1	#10	#11	#12
MLP2	#13	#14	#15
MLP3	#16	#17	#18

将算法用集合 alg 表示,特征用集合 fea 表示,alg 与 fea 做笛卡儿积,则有 12 个有序对组合(表 3 的#1 ~ #12),用一个数字编号表示一种组合,例如,#1 表示(API,LR),代表使用 API 特征和 LR 算法训练的一个分类器。考虑到基于 MLP 的模型根据隐层数不同可能存在较大的性质差异,对#10 ~ #12 分类器额外设置了 2 组不同层数的分类器,因此共有 18 个目标分类器。

3.4 实验设计

攻击实验中,恶意样本数据集有 60% 用于目标分类器的训练,在剩余的 40% 中,在 9 个恶意代码家族各随机选取了 20 个样本,将这 180 个样本称为攻击样本,用于生成对应的对抗样本。

使用本文的对抗样本生成方法,生成上述 180 个攻击样本的对抗本来攻击每一个目标分类器。用扰动常量控制扰动大小^[25],绘制扰动大小-TPR 图、扰动大小-ASR 图,以获得不同扰动强度下的攻击效果。为进一步增强说服力,本文设计了对比实验:复现 MalGAN 和 ZOO 这 2 个先进的黑盒对抗样本生成方法,生成攻击样本的对抗样本,攻击#1、#4、#7、#10、#13、#16 分类器,记录 TPR 攻击前后的变化,并与本文方法进行对比分析。

3.5 实验与结果

图 2 为基于不同扰动维度(Dr)生成对抗样本,攻击各目标分类器产生的 TPR 和 ASR 的变化图。查看 TPR 图可以发现,18 个分类器的 TPR 均降到了接近 0 的水平,即在较高扰动代价情况下,本文方法几乎 100% 成功地攻击任意恶意代码分类器。

对比实验中,使用 MalGAN 和 ZOO 生成攻击样本的对抗样本,攻击#1、#4、#7、#10、#13、#16 分类器。每个分类器性质不同,选用参数也有差异,本文倾向于选择使对抗样本更有效的参数,经过多次实验,结合本文方法汇总如表 4 所示。

从表 4 可以看出,MalGAN 和本文方法都具有较好的效果,且攻击效果相似(将 TPR 降到更

低,且降幅相似)。ZOO 在实验中表现不佳,攻击后的 TPR 都在 50% 以上;MalGAN 将#1、#10、#13、#16 分类器的 TPR 降到 0, #4、#7 分类器的 TPR 降到 1.67% 和 0.56%;本文方法将#1、#4、#10、#16 分类器的 TPR 降到 0,而#7、#13 分类器的 TPR 分别降到 1.67% 和 1.11%。为了进一步对比这 3 种方法,比较了 3 种方法生成的对抗样

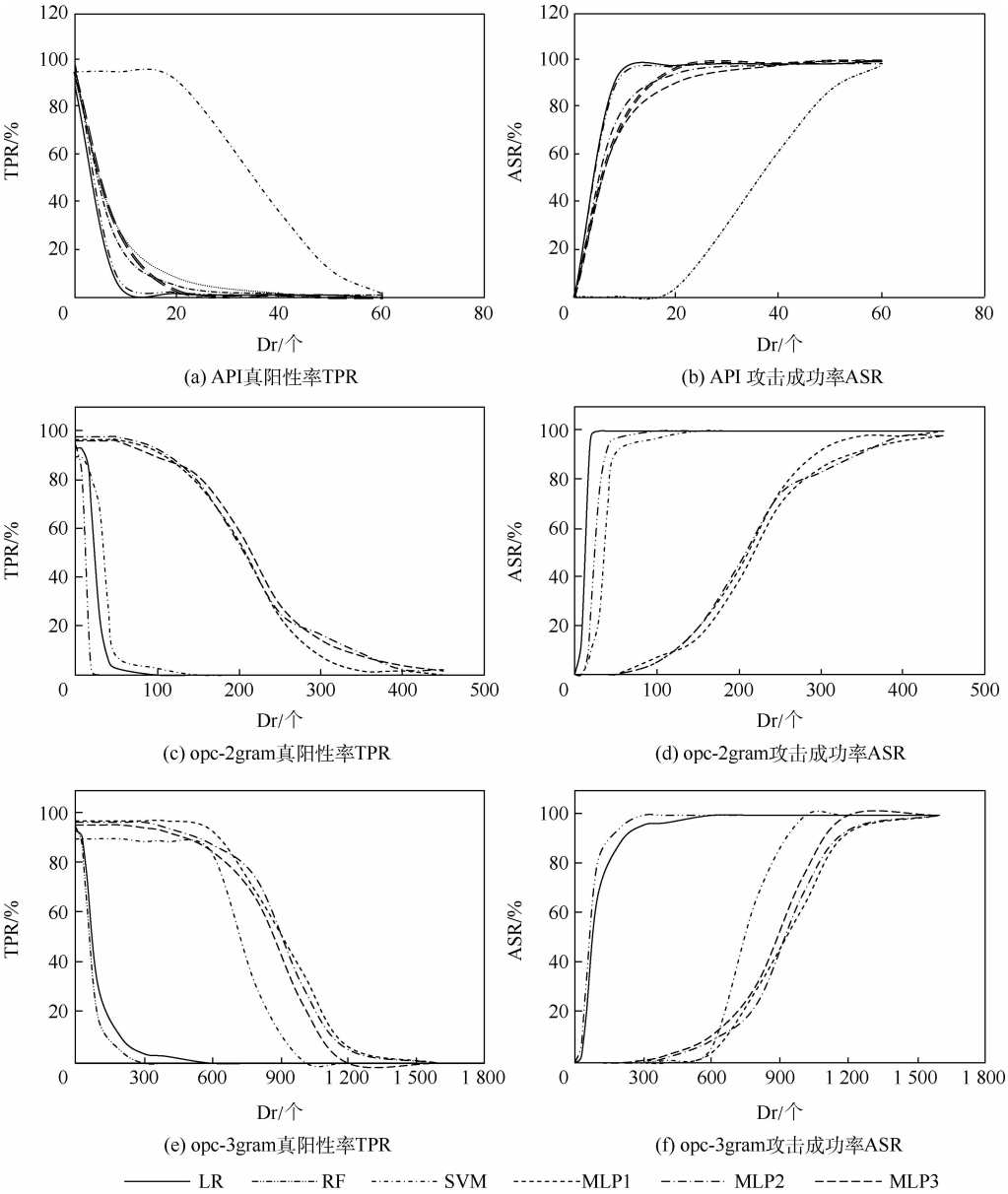


图 2 三种特征的 Dr-TPR 图和 Dr-ASR 图

Fig. 2 Dr-TPR and Dr-ASR of three characteristics

表 4 API 特征分类器的真阳性率对比

Table 4 API feature classifier and TPR comparison

对抗样本 生成方法	TPR/%					
	#1-LR	#4-RF	#7-SVM	#10-MLP1	#13-MLP2	#16-MLP3
无攻击对照	89.44	98.89	95.38	92.22	95.56	95.00
ZOO	57.78	91.11	61.67	69.44	72.78	70.56
MalGAN	0	1.67	0.56	0	0	0
本文方法	0	0	1.67	0	1.11	0

本本身的差异。

实验中的 API 特征是 One-Hot 型的,对应取值应该是 0 或者 1,而 ZOO 在特征中可能出现 -3、-1、2 等值,只能通过筛选来获取符合要求的对抗样本,这将导致有效对抗样本进一步减少;MalGAN 适用于 One-Hot 类型的特征,主要对比其

生成的对抗样本的扰动维度。表 5 为 MalGAN 和本文方法在对各分类器取得较好攻击效果时的平均扰动维度 D_r 。其中,攻击效果相似的情况下,本文方法比 MalGAN 生成的样本平均扰动维度小,这可能因为本文方法有针对扰动大小的设计。

表 5 两种方法生成的对抗样本平均扰动维度

对抗样本 生成方法	Dr/个					
	#1-LR	#4-RF	#7-SVM	#10-MLP1	#13-MLP2	#16-MLP3
MalGAN	28.31	23.26	102.87	24.70	35.14	29.45
本文方法	9.56	11.62	43.12	15.47	13.39	13.92

4 结 论

1) 本文方法是一种有效的黑盒对抗样本生成方法。使用该方法生成的对抗样本测试黑盒的恶意代码分类器,能显著降低分类器的真阳性率。

2) 本文方法适用范围广泛。使用该方法攻击 18 个不同算法或特征的目标分类器,均有不错的攻击成功率。目标分类器的算法涵盖了线性算法、树形算法和深度神经网络算法,而特征既有 One-Hot 型也有数值型。

3) 本文方法能有效控制扰动的大小。可以通过设置不同的扰动常量来控制对抗样本的扰动维数和扰动范围。

4) 本文方法具有健全性。随着扰动的增大,攻击成功率是严格递增的;随着扰动持续增大,攻击成功率接近或达到 100%。虽然过大的扰动可能会使对抗样本失去意义,但可以说明该方法是健全的。

参考文献 (References)

[1] ALAZAB M. Automated malware detection in mobile app stores based on robust feature generation [J]. Electronics, 2020, 9 (3):435.

[2] SAXE J, BERLIN K. Deep neural network based malware detection using two dimensional binary program features [C] // 2015 10th International Conference on Malicious and Unwanted Software (MALWARE). Piscataway: IEEE Press, 2015: 11-20.

[3] PASCANU R, STOKES J W, SANOSSIAN H, et al. Malware classification with recurrent networks [C] // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2015: 1916-1920.

[4] HUANG W Y, STOKES J W. MtNet: A multi-task neural network for dynamic malware classification [C] // International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA). Berlin: Springer, 2016: 399-418.

[5] KOLOSINAJI B, ZARRAS A, WEBSTER G, et al. Deep learning for classification of malware system call sequences [C] // Australasian Joint Conference on Artificial Intelligence. Berlin: Springer, 2016: 137-149.

[6] SCHULTZ M G, ESKIN E, ZADOK F, et al. Data mining methods for detection of new malicious executables [C] // Proceedings 2001 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press, 2000: 38-49.

[7] KOLTER J Z, MALOOF M A. Learning to detect malicious executables in the wild [C] // Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 470-478.

[8] KOLTER J Z, MALOOF M A. Learning to detect and classify malicious executables in the wild [J]. Journal of Machine Learning Research, 2006, 7 (4): 2721-2744.

[9] RIBEIRO M T, SINGH S, GUESTRIN C. “ Why should I trust You? ”: Explaining the predictions of any classifier [C] // Proceedings of the 2016 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1135-1144.

[10] SU D, ZHANG H, CHEN H G, et al. Is robustness the cost of accuracy? A comprehensive study on the robustness of 18 deep image classification models [C] // Computer Vision-ECCV 2018, 2018: 644-661.

[11] STOKES J W, WANG D, MARINESCU M, et al. Attack and defense of dynamic analysis-based, adversarial neural malware detection models [C] // 2018 IEEE Military Communications Conference (MILCOM). Piscataway: IEEE Press, 2018: 1-8.

[12] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C] // Proceeding of 2nd International Conference on Learning Representations (ICLR), 2014: 14-16.

[13] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [C] // Proceedings of 3rd International Conference on Learning Representations (ICLR), 2015: 7-9.

[14] HU W, TAN Y. Generating adversarial malware examples for black-box attacks based on GAN [EB/OL]. (2017-02-20) [2020-08-01]. https://arxiv.org/abs/1702.05983.

[15] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: Zeroth order

optimization based black-box attacks to deep neural networks without training substitute models[C] // Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2017: 15-26.

[16] DOBROVOLJC A, TREEK D, LIKAR B. Predicting exploitations of information systems vulnerabilities through attackers' characteristics[J]. IEEE Access, 2017, 5: 26063-26075.

[17] CRANDALL J R, SU Z D, WU S F. On deriving unknown vulnerabilities from zero-day polymorphic and metamorphic worm exploits[C] // Proceedings of the 12th ACM Conference on Computer and Communications Security. New York: ACM, 2005: 235-248.

[18] ÍNCER R Í, THEODORIDES M, AFROZ S, et al. Adversarially robust malware detection using monotonic classification[C] // Proceedings of the 4th ACM International Workshop on Security and Privacy Analytics. New York: ACM, 2018: 54-63.

[19] XU W L, EVANS D, QI Y J. Feature squeezing: Detecting adversarial examples in deep neural networks[EB/OL]. (2017-12-05) [2020-08-01]. <https://arxiv.org/abs/1704.01155>.

[20] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C] // 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.

[21] KREUK F, BARAK A, AVIV-REUVEN S, et al. Adversarial examples on discrete sequences for beating whole-binary malware detection[EB/OL]. (2019-01-10) [2020-08-01]. <https://arxiv.org/abs/1802.04528v1>.

[22] SWIESKOWSKI P, KUZINS S. Ninite-install or update multiple apps at once[EB/OL]. (2015-10-01) [2020-08-01]. <https://ninite.com/>.

[23] RAFF E, BARKER J, SYLVESTER J, et al. Malware detection by eating a whole exe[C] // The Workshops of the Thirty-Second Conference on Artificial Intelligence (AAAI), 2018: 268-276.

[24] GOLDBLOOM A, HAMNER B. Microsoft malware classification challenge (BIG 2015) [EB/OL]. (2015-12-15) [2020-08-01]. <https://www.kaggle.com/c/malware-classification/data>.

[25] CARLINI N, ATHALYE A, PAPERNOT N, et al. On evaluating adversarial robustness [EB/OL]. (2019-02-20) [2020-08-01]. <https://arxiv.org/abs/1902.06705>.

Adversarial sample generation technology of malicious code based on LIME

HUANG Tianbo, LI Chengyang, LIU Yongzhi, LI Denghui, WEN Weiping*

(School of Software & Microelectronics, Peking University, Beijing 102600, China)

Abstract: Based on the research and analysis of machine learning technology to detect malicious code, a local interpretable model-agnostic explanations (LIME)-based black-box adversarial examples generation method is proposed to generate adversarial samples for any black-box malicious code classifier and bypass the detection of machine learning models. The method uses a simple model to simulate the target classifier's local performances, obtains the feature weights, and generates disturbances through the disturbance algorithm. According to the generated disturbances, the method modifies the original malicious code to generate adversarial samples. We test the method using Microsoft's common malicious sample data in 2015 and the collected benign sample data from more than 50 suppliers as follows: 18 target classifiers based on different algorithms or features were implemented concerning common malicious code classifiers. Their classifiers' true positive rates were reduced to approximately zero when we attacked them using the method. Two advanced black-box sample generation methods, MalGAN and ZOO, were reproduced for comparison with this method. The experimental results show that the proposed method in this paper can effectively generate adversarial samples, and the method itself owns various strengths, including broad applicability, flexible control of disturbances, and soundness.

Keywords: adversarial samples; malicious code; machine learning; local interpretable model-agnostic explanations (LIME); target classifiers

Received: 2020-08-09; **Accepted:** 2020-10-23; **Published online:** 2021-01-21 09:57
URL: kns.cnki.net/kcms/detail/11.2625.V.20210121.0934.001.html
Foundation item: National Natural Science Foundation of China (61872011)
*** Corresponding author.** E-mail: weipingwen@ss.pku.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0529

一种基于攻击距离的对抗样本攻击组筛选方法

刘洪毅, 方字形, 文伟平*
(北京大学 软件与微电子学院, 北京 102600)

摘 要: 黑盒对抗样本生成过程中通常会指定 1 个攻击组, 包括 1 个原始样本和 1 个目标样本, 使得生成的对抗样本与原始样本范数差别不大, 但被分类器识别为目标样本的分类。针对攻击组的攻击难度不同导致攻击不稳定的问题, 以图像识别领域为例, 设计了基于决策边界长度的攻击距离度量方法, 为攻击组的攻击难易程度提供了度量方法。在此基础上, 设计了基于攻击距离的对抗样本攻击组筛选方法, 在攻击开始前就筛去难以攻击的攻击组, 从而实现在不修改攻击算法的前提下, 提升攻击效果。实验表明: 相比于筛选前的攻击组, 筛选后的攻击组的总体效果提升了 42.07%, 攻击效率提升了 24.99%, 方差降低了 76.23%。利用攻击组的对抗样本生成方法在攻击前先进行攻击组筛选, 可以稳定并提高攻击效果。

关 键 词: 对抗样本; 黑盒; 决策边界; 筛选; 图像识别
中图分类号: TP319
文献标志码: A **文章编号:** 1001-5965(2022)02-0339-09

随着深度学习算法及计算设备的快速发展, 深度学习由于其优秀的效果与性能被广泛应用于各个领域, 如语音识别、图像识别、人脸识别、自动驾驶等。在计算机视觉领域, 随着 2012 年 Krizhevsky 等^[1]设计 AlexNet 赢得了 ImageNet LS-VRC-2012 比赛的冠军, 奠定了卷积神经网络在深度学习应用中的地位。在之后的发展中, 学者们先后提出了 GoogLeNet^[2]、VGG^[3]、ResNet^[4]、DenseNet^[5]等深层网络模型, 使得基于深度学习的图像识别技术成为了图像识别领域的主流, 并使得识别精准度在 ImageNet 上达到了 95% 以上。在带来便利的同时, 深度学习本身也存在着一定的安全问题, 这引起了安全领域的极大关注。例如, Szegedy 等^[6]很快发现了深层神经网络在图像识别领域的一个有趣的弱点, 即深度神经网络容易被对抗样本欺骗。这种攻击的表现形式是对图像添加一个微小的扰动, 并且这些扰动对于人类

视觉来讲, 是几乎不可察觉的, 但会使得分类器分类错误。

随着研究的深入, 学者们将攻击方法进行了分类。根据攻击者掌握的知识, 对抗攻击可分为白盒攻击和黑盒攻击。其中, 白盒攻击是指攻击者完全了解模型的架构和参数。白盒模型在实际的威胁模型中并不常见, 更为广泛的应用场景是黑盒模型。因此, 更关注黑盒环境下的对抗样本生成技术。黑盒攻击指攻击者无法获取模型全部信息, 只能通过对模型的使用来观察输入输出并展开攻击。

黑盒攻击方法主要有 2 种, 基于可转移性的对抗样本生成方法 (transfer-based adversarial example generation, TAEG) 和基于查询的对抗样本生成方法 (query-based adversarial example generation, QAEG)。可转移性是 对抗样本的一种性质, 即针对某分类器模型生成的对抗样本也可能欺骗

收稿日期: 2020-09-21; 录用日期: 2020-11-06; 网络出版时间: 2021-01-20 08:56
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210119.1721.001.html
基金项目: 国家自然科学基金 (U1736218)
* 通信作者. E-mail: weipingwen@ss.pku.edu.cn

引用格式: 刘洪毅, 方字形, 文伟平. 一种基于攻击距离的对抗样本攻击组筛选方法[J]. 北京航空航天大学学报, 2022, 48(2): 339-347. LIU H Y, FANG Y T, WEN W P. A method for filtering the attack pairs of adversarial examples based on attack distance [J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 339-347 (in Chinese).

其他分类器模型。TAEG 方法主要利用对抗样本的可转移性,针对一个替代模型进行白盒攻击,如基于梯度的对抗样本生成方法,包括 FGSM^[7]、I-FGSM^[8]、PGD^[9]、MI-FGSM^[10]、CW^[11]、JSMA^[12]、DeepFool^[13]等,并使用产生的对抗样本攻击目标模型,如 Substitute Model^[14]、DI-2-FGSM^[15]、Ensemble Attack^[16]。基于可转移性的对抗攻击可以仅查询目标模型一次就完成攻击,但其攻击成功率较低,若非目标模型和替代模型十分相似,即便是目前鲁棒性最好的方法,攻击成功率也难以令人满意^[15]。而 QAEG 方法已经被实践^[17-18]证明可以应用在现实模型上。QAEG 仅需要对目标模型进行一定的查询,根据查询结果不断优化对抗样本,即可实现黑盒攻击,并且成功率一般都较高。

目前流行的 QAEG 方法都是基于攻击组的,如 NES Attack^[17]、CMA Attack^[18-19]、Boundary Attack^[20]等。这些方法引入原始样本与目标样本(称这一对样本为一个攻击组)共同参与攻击,当对抗样本生成任务陷入局部最优解时,为对抗样本进化提供了基本的方向,即向着原始样本靠近,可以在一定程度上缓解由于估算梯度不准确而导致的对抗样本陷入局部最优的情况。但是,由于引入了攻击组来指导攻击,针对一个对抗样本生成任务,如何合理地选择对抗攻击组将成为一个挑战。由于对抗攻击的目的仅是使得目标模型错误分类。但在现有的研究中,由于引入了攻击组的指导,相当于加入了一个限制条件,当生成任务到达局部最优解时,只能向着原始样本靠近。但原始样本与目标样本可能并不容易实现攻击,最终导致攻击失败,或者所需查询次数过多,如 CMA Attack 实现攻击的平均查询次数为 60 000 次,但最坏的情况可能达到 200 000 次^[18]。

为了解决以上挑战,本文以图像识别领域为例,设计了一种基于决策边界长度的攻击距离度量方法,探测原始图片到目标分类决策边界的垂直距离,并计算需要进化的决策边界长度,以度量两图片的攻击距离,并为该攻击组的攻击难易程度提供参考。同时,设计了基于攻击距离的对抗样本攻击组筛选方法,利用图片间的攻击距离,选取较为容易完成攻击的攻击组,再展开对抗样本生成任务。通过实验验证,筛选后的攻击组对于多个攻击方法有效,总体攻击效果提升 42.07%,攻击效率提升 24.99%,方差降低 76.23%。

1 相关工作

1.1 深度学习

深度学习^[21]是机器学习领域的一个新的研究方向,其目前正以前所未有的规模被用于破解各种复杂的科学难题。例如,深度神经网络(deep neural networks, DNN)在图像识别、重建脑回路^[22]、DNA 突变分析^[23]等各种任务上都取得了显著的成功。DNN 也成为语音识别^[24]、自然语言理解^[25]、无人机^[26]、机器人^[27]和人脸识别 ATM^[28]中许多具有挑战性任务的首选解决方案。显然,深度学习解决方案,尤其是源自计算机视觉问题的解决方案,将在日常生活中发挥重要作用。

1.2 对抗样本

对于深度学习分类器来说,对抗样本是一种特殊的样本,其与原始样本的差别不大,但却可以使得分类器分类错误。对抗样本生成的基本思想是:对于原始图片 x ,设计一个尽量小的扰动 δ ,能够使得分类器 C 分类错误,如式(1)所示,上述场景被称为非目标攻击,即只要使得分类器分类错误即可。

$$\min \|\delta\|$$
$$\text{s. t. } C(x + \delta) \neq C(x) \tag{1}$$

同时,还有一种更强的攻击场景,要求对抗样本使得分类器分类为指定标签 t ,如下:

$$\min \|\delta\|$$
$$\text{s. t. } C(x + \delta) = t \tag{2}$$

其中扰动大小的度量公式常使用 P 范式,如式(3)所示。本文中使用 L_2 距离度量扰动的大小,即 2 范式、欧氏距离,其中 N 为像素点的个数。为了能够在连续的空间上搜索对抗样本,将像素值从 $[0, 255]$ 压缩到 $[0, 1]$ 。同时,给出攻击成功的判定,一般为平均像素差值小于 0.05 且使分类器分类错误,对应到大小为 $(299, 299, 3)$ 的图片上, L_2 距离约为 25。

$$\|x\|_p = \left(\sum_{i=1}^N |x_i|^p \right)^{\frac{1}{p}} \tag{3}$$

白盒对抗攻击最常用的就算是基于梯度的对抗样本生成方法,通过计算损失函数对输入图片的梯度,再通过梯度下降的方式降低损失值以达成目的。其中损失函数设计原理为损失值越小,对抗样本的效果越好,如扰动的 L_2 距离加上目标分类置信度的负值。Szegedy 等^[6]首先设计了让神经网络做出误分类的最小扰动的方程,转而寻找最小的损失函数添加项,将对抗样本生成问题

转化成了最优化问题。Goodfellow 等^[7]使用 L_{inf} 距离设定扰动的上限,将原始分类的交叉熵作为目标函数,最大化目标函数以实现非目标攻击。Kurakin^[8]和 Makelov^[9]等引入了迭代的思想,将扰动上限分成 n 份逐步完成攻击。Carlini 和 Wagner^[11]将错误分类的约束条件统一至目标函数中,将对抗样本生成问题转化为优化问题,并设计了强有力的攻击方法。Papernot 等^[12]提出使用梯度显著图的方式,每轮迭代仅选择效果最优的 2 个像素点进行修改,尽量降低扰动大小的同时完成了攻击。

1.3 黑盒对抗攻击

黑盒对抗攻击现有 2 种常用的方法,即 TAEG 和 QAEG。

可转移性是对抗样本的一个很重要的性质。可转移性是指针对某一个模型产生的对抗样本,也可以欺骗其他模型^[29]。对抗样本存在可转移性的主要原因是:同类型分类任务学到的分类区域可能大致相同,这导致对于一个模型有效的对抗样本可能对另一模型也有效。Papernot 等^[14]通过一定量的查询来构造与目标模型相似的替代模型,再使用白盒方法攻击这个替代模型以产生能够攻击目标黑盒模型的对抗样本。Dong 等^[10]将动量引入对抗样本生成过程,使得梯度更新的方向更加稳定,提高了攻击鲁棒性,同时提高了对抗样本的可转移性。Xie 等^[15]将输入变换引入到对抗样本生成过程,提高了对抗样本的可转移性。Tramer 等^[16]提出了联合多个模型进行生成及防御对抗样本,大幅提高了对抗样本的可转移性。Huang 等^[30]提出了针对模型的中间层而非 logit 层或输出层进行攻击,可以一定程度上提高对抗样本的可转移性。

尽管有众多学者对对抗样本的可转移性进行了研究,但是 TAEG 的黑盒成功率依然很低^[15],难以适应实际的需求。因此,QAEG 受到了越来越多的关注。QAEG 是通过查询目标模型,来估算模型对于输入的自然梯度,并辅以原始样本与目标图样本,逐步地进化对抗样本。在目标攻击场景下,QAEG 最常用基于攻击组的对抗攻击方法,即从目标图片出发,保持对抗样本的分类为目标分类,持续地降低对抗样本到原始图片的距离。Ilyas 等^[17]提出使用自然进化策略(natural evolutionary strategies, NES^[31])来估算目标模型输出对抗样本分布期望的梯度,并通过进化整个分布来进化对抗样本。Kuang 等^[18]提出使用协方差矩阵自适应进化策略(covariance matrix adaptation

evolutionary strategies, CMA-ES^[32])来拟合并进化对抗样本的分布,提升查询模型的效率,以降低完成攻击所需的查询次数。Brendel 等^[20]提出通过拒绝采样来拟合决策边界,不断地进化对抗样本,削弱了攻击算法对于模型自然梯度的依赖性,并实现了仅利用模型输出的 top1 标签完成攻击。Dong 等^[19]降低了搜索维度,并利用 $(1+1)$ -CMA-ES 来拟合决策边界分布,快速地进化对抗样本以实现攻击。

2 方法论

2.1 威胁模型

目前,攻击组的引入主要是为了解决黑盒模型下,目标攻击估算的自然梯度不精确的问题。因此,本文主要研究黑盒模型下目标攻击的攻击组筛选方法,同时要求该黑盒模型可以被查询多次。

现有目标性对抗攻击主要有如下 4 种应用场景:

- 1) 场景 1。原始分类固定,目标分类固定,但原始图片与目标图片均不固定。这是限定最弱的攻击场景,也是最常见的攻击场景,仅实现目标性攻击即可。实现人类识别结果和机器识别结果不同,如仅使目标分类系统出现错误。
- 2) 场景 2。原始图像固定,目标分类固定,但目标图像不固定。这是另一种常用的攻击场景,给定原始图像,使得对抗样本与原始图像相差不大,但被分类器分类为目标分类,如自动驾驶欺骗。
- 3) 场景 3。原始分类固定,目标图片固定,但原始图片不固定。这也是一种常见的攻击场景,如绕过人脸身份验证系统。
- 4) 场景 4。原始图像固定,目标图像固定。这是最严格的攻击场景,但并不常见。

在前 3 种应用场景中,都存在很多符合要求的攻击组。本文主要研究如何在前 3 种场景中,筛选出较为容易的攻击组,从而提高现有黑盒攻击的效率及攻击的稳定性。

2.2 基于攻击距离的攻击组筛选

现有的 QAEG 虽然普遍使用了攻击组来指导攻击,但未注意到由于攻击组难度差异导致的攻击效果不稳定问题。因此,目前也没有估算攻击组攻击难度的方法。为了解决以上问题,本文提出了基于决策边界长度的攻击距离度量方法,讨论了如何利用估算的攻击距离筛选对抗样本攻击组,通过预先筛选现有对抗样本生成方法的输入来大幅提升其攻击效率。基于攻击距离的对抗样本攻击组筛选方法的框架如图 1 所示。

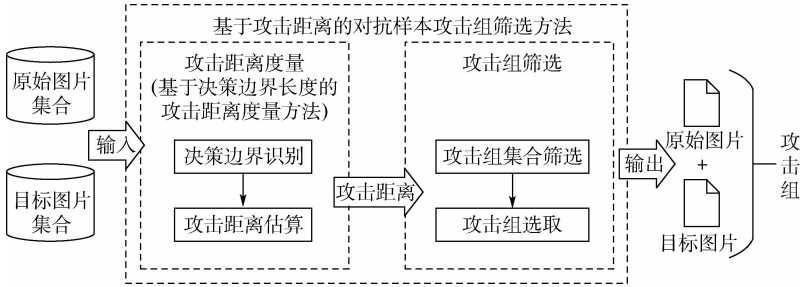


图 1 基于攻击距离的对抗样本攻击组筛选方法框架

Fig. 1 Framework of adversarial example attack pairs filtering method based on attack distance

2.2.1 攻击距离度量

以 NES Attack 为例,其对抗样本生成的流程如图 2 所示,攻击速度如图 3(a)所示。图 2 中: s_img 为原始图片, t_img 为目标图片, adv_i 为对抗样本的第 i 轮迭代结果。迭代的主要内容就是通过查询目标模型,估算目标函数对输入的梯度,并通过梯度下降的方式更新对抗样本。同时,在攻击速度图中可以明显看到,对抗样本的 L_2 距离会先快速下降,再逐渐平缓。CMA Attack 的攻击速度也是类似的,如图 3(b)所示。这其中的原因是:基于查询的攻击都会先快速找到决策边界,这个过程一般很快,再通过迭代稳步地更新对抗样本,这个过程一般很慢。而重要的是:第 2 个过程

可以用线性拟合,以提前估算对抗样本生成所需要的查询次数。即查询次数和扰动的大小成反比,如果能够提前估算第 2 个阶段完成的攻击距离,就可以提前估算完成攻击所需的查询次数,从而估算攻击组的攻击难度。

图 3 展示了 NES Attack^[17] 与 CMA Attack^[18] 在 ImageNet 上攻击 InceptionV3 的攻击速度。2 个

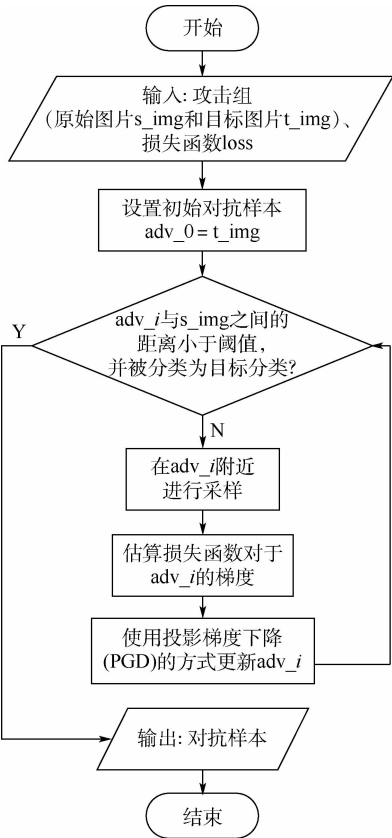


图 2 NES Attack 对抗样本生成流程

Fig. 2 Flowchart of NES Attack adversarial examples generation

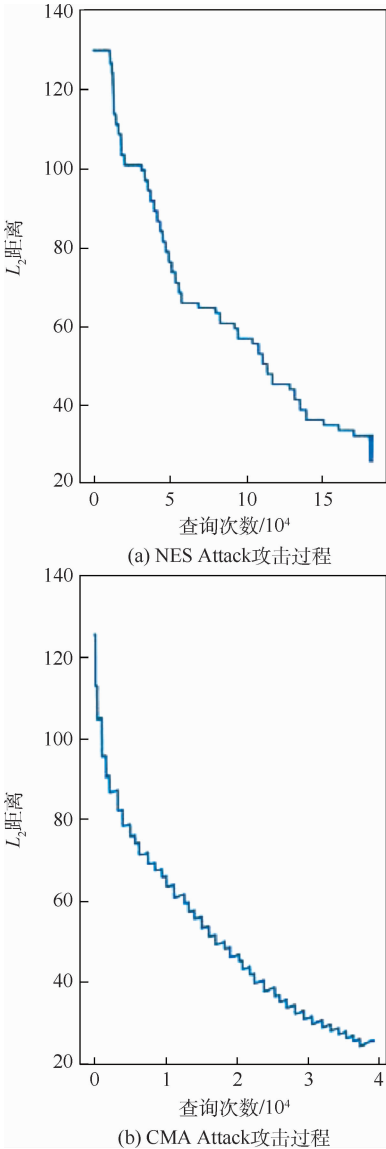


图 3 攻击速度

Fig. 3 Attack speed

方法针对同一攻击组展开攻击,并展现出相似的攻击速度曲线。

基于查询的攻击过程有上述特性主要是由于对抗样本的特性。对抗样本是一种既要被分类为错误分类,又要与原始图片相差不大的样本。这就要求对抗样本一般都处于目标分类与原始分类的分类边界附近。因此,对抗样本迭代生成的过程就是对抗样本游走于决策边界,并一步步向原始图片靠近的过程,包括基于梯度的白盒攻击与基于查询的黑盒攻击。估算一个攻击组完成攻击所需要的查询次数,主要需要解决 2 个问题:①如何确定对抗样本已经抵达决策边界附近;②如何估算攻击距离。

为了解决第 1 个问题,设计了基于采样的决策边界识别方法,其效果示意图如图 4 所示。将原始图片与目标图片设为 2 个端点,并在 2 端点所在连线上进行二分查找,以快速定位决策边界位置。查找规则如下:如果大部分采样均为目标分类,则将中间图片向原始图片靠近;如果大部分采样均为原始分类,则将中间图片向目标图片靠近;直到在中间图片附近的采样有 40% ~ 60% 的概率为目标分类。图 4(a)展示了中间图片附近采样分类为目标分类的概率小于 40% 的情况,中间图片将向目标图片靠近,即图 4(b)展现的情况。

为了解决第 2 个问题,设计了基于毕达哥拉斯定理的攻击距离估算方法,其效果示意图如图 5 所示。设中间图片到原始图片的距离为 L_m ,最终对抗样本到原始图片的距离为 L_d ,从中间图片进化到对抗样本需要完成的攻击距离为 L_a 。由于对抗样本进化的过程实际是在决策边界附近游走的过程,原始图片到最终对抗样本的距离可以看作是原始图片到决策边界的距离。又因为点到面垂线距离最短,所以最先找到符合 L_d 距离要求的对抗样本应该处于原始图片到决策边界的垂线附近。同时,对抗样本游走的攻击路径在决策边界附近,因此可以看做 L_d 近似垂直于 L_a 。因此由毕达哥拉斯定理得, $L_a^2 = L_m^2 - L_d^2$ 。整个基于决策边界长度的攻击距离度量方法如算法 1 所示。

算法 1 基于决策边界长度的攻击距离度量方法。

输入:原始图片 s_img 、目标图片 t_img 、目标模型 $C()$ 、期望的对抗样本的扰动大小 L_d 。

输出:攻击距离 L_a 。

```
1: mid = t_img, low = s_img, high = t_img
2: do:
```

```
3:   $X \sim N(0,1)$ 
4:  在 mid 附近生成采样  $X' = mid + X \times 2.55$ 
5:  利用  $C()$  计算  $X'$  中对抗样本的比例  $P$ 
6:  IF  $P \geq 60\%$  :
7:      mid = (mid + low) / 2
8:      high = mid
9:  ELSE IF  $P \leq 40\%$  :
10:     mid = (mid + high) / 2
11:     low = mid
12:  END IF
13: WHILE  $P \geq 60\% \parallel P \leq 40\%$ 
14:  $L_m = \|mid - s\_img\|_2$ 
15:  $L_a = \sqrt{L_m^2 - L_d^2}$ 
16: RETURN  $L_a$ 
```

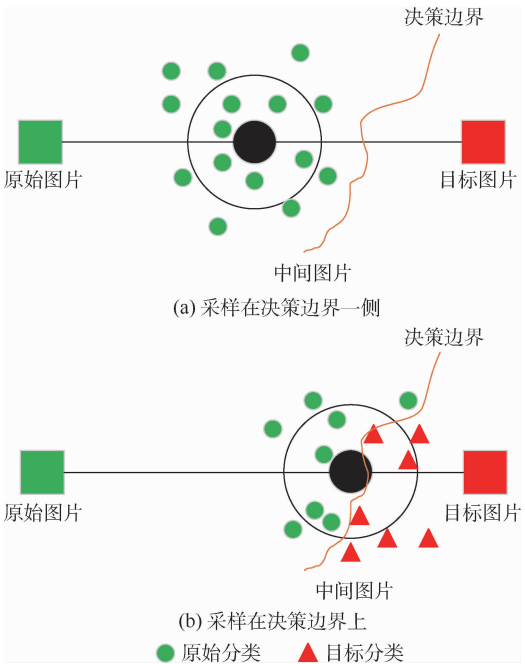


图 4 基于采样的决策边界识别方法示意图

Fig. 4 Schematic diagram of decision boundary recognition method based on sampling

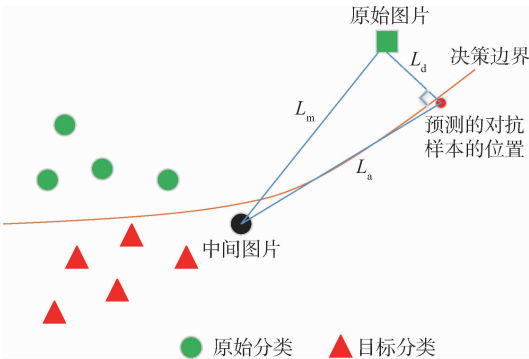


图 5 攻击距离估算示意图

Fig. 5 Schematic diagram of attack distance estimation

2.2.2 攻击组筛选

估算出攻击组之间的攻击距离之后,就可以

在现有攻击方法之前做一个预处理阶段,进行攻击组的筛选,以提高现有算法的攻击效果。

需要注意的是,攻击距离的估算结果并非十分精确,能大致反映完成攻击所需要的查询次数,但也会存在一些偏差。因此,不能直接利用估算的攻击距离选出最容易攻击的攻击组,但可以利用估算的攻击距离筛选出较容易攻击的攻击组。具体来说,从攻击组集合 S 中筛选出较容易攻击的攻击组集合 E ,再从 E 中随机选取一组作为建议展开攻击的攻击组 Pair。三者之间的关系如下:

$$\text{Pair} \in E \quad E \subset S \tag{4}$$

其中,将集合 E 与集合 S 的比例定为超参数,即过滤因子 R 。关于 R 值对于算法的影响,将在第 3 节进行详细的讨论。筛选时会先利用算法 1 估算 S 中各个攻击组的攻击距离,并将攻击组按照攻击距离升序排序。之后筛选前 $\eta = \rho R$ 个攻击组来组成 E , ρ 为 S 中攻击组的个数, η 为集合 E 中攻击组的个数。最终在 E 中随机选取一组作为 Pair。基于攻击距离的对抗样本攻击组筛选方法如算法 2 所示。

算法 2 基于攻击距离的对抗样本攻击组筛选方法。

输入:原始图片集合 s_imgs 、目标图片集合 t_imgs 、过滤因子 R 。

输出:合适的攻击组 Pair。

1: 生成攻击组集合

$$S = \{(s_img, t_img) \mid s_img \in s_imgs, t_img \in t_imgs\}$$

2: $\rho = \text{len}(S)$

3: 利用算法 1 获取 S 中每一组的攻击距离

4: 将攻击组按照攻击距离降序排序

5: $\eta = \rho R$

6: 选取前 η 个攻击组组成 E

7: $\text{index} = \text{rand}(0, \eta)$

8: $\text{Pair} = E[\text{index}]$

9: RETURN Pair

3 实验与评估

本节将讨论攻击组筛选对不同的攻击方法在不同应用场景下的有效性,同时讨论超参数 R 对于算法的影响。

本文主要对比的攻击方法包括 NES Attack^[17]和 CMA Attack^[18],目标模型为 InceptionV3。对比攻击组筛选前后实现攻击所需要查询次数的平均

值、中位数及方差,从而验证攻击组筛选的有效性。

实验设定如下:选取 ImageNette 数据集进行验证与评估,共 10 个分类,场景 1 每种分类选取 10 张图片参与攻击组的构造,场景 2、3 每种分类选取 100 张图片。场景 1 与场景 2、3 实验环境设置不同是因为场景 1 的组合情况过多,若每种分类选取图片均为 100 张,场景 1 的攻击组个数会达到 $900\,000 = 90 \times 100 \times 100$,而场景 2、3 的攻击组个数仅为 $9\,000 = 90 \times 100$ 。因此,降低场景 1 每种分类选取的图片为 10 张,以平衡 3 种场景的攻击组个数均为 $9\,000 = 90 \times 10 \times 10$ 。

具体来说,从 10 个分类中选择原始分类和目标分类,共有 90 种不重复的组合。场景 1 时,分类 A 中从 10 张图片选取 1 张作为原始图片,分类 B 中也从 10 张图片选取 1 张作为目标图片,共有 100 种攻击组合,因此场景 1 共有 9 000 种攻击组。场景 2 时,分类 A 中固定选取 1 张图片作为原始图片,分类 B 中从 100 张图片中选取 1 张作为目标图片,共有 100 种攻击组合,因此场景 2 也共有 9 000 种攻击组。场景 3 与场景 2 类似,分类 A 中从 100 张图片选取 1 张作为原始图片,分类 B 中固定选取 1 张图片作为目标图片,共 100 种组合,因此场景 3 也共有 9 000 种攻击组。

本文首先讨论不同 R 值在场景 1 下对不同攻击算法的影响,如图 6 所示。随着 R 值的降低,2 种方法的平均查询次数与查询次数中位数均稳定地下降,验证了攻击组筛选对于提高查询效率的有效性。同时随着 R 值的降低,CMA Attack 的查询次数方差稳定下降,NES Attack 的查询次数方差波动下降,这验证了攻击组筛选对于稳定方差的有效性。实验结果显示, R 值越小,攻击效果越好,因此推荐将 R 值设为 0.1。本文之后的性能对比也是基于 R 值为 0.1 的基础上进行的。

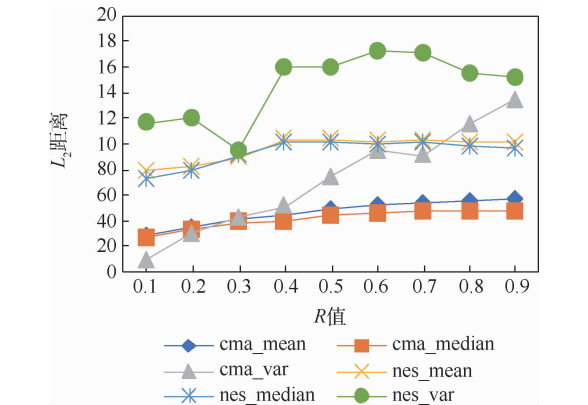


图 6 R 值影响

Fig.6 Influence diagram of R 's value

图 6 中: cma_mean、cma_median 和 cma_var 分别代表 CMA Attack 攻击成功所需查询次数的平均值、中位数和方差; nes_mean、nes_median 和 nes_var 分别代表 NES Attack 攻击成功所需要查询次数的平均值、中位数和方差;同时为了共用纵坐标轴,方差曲线显示的值为实际方差的 1/10 000。

攻击组筛选前后的攻击效果对比如表 1 所示。表 1 展现了不同对抗样本生成方法在不同场景下实现攻击所需要的查询次数的平均值、中位数和方差。表 1 给出了所有待测的方法,包括 NES Attack^[17]、CMA Attack^[18]、经过攻击组筛选的 NES Attack (Filtered NES) 和经过攻击组筛选的 CMA Attack (Filtered CMA)。

表 1 中给出了攻击组筛选在每一个指标上带来的平均提升效果。总体提升效果表示攻击组筛选在不同场景下平均值、中位数和方差的综合提升,即将所有平均提升效果求平均;平均提升效率表示攻击组筛选在不同场景下的平均值、中位数的综合提升,即将不同场景下的平均值与中位数提升效果统一求平均;平均稳定方差表示攻击组筛选在不用场景下的平均稳定方差效果,即将不同场景下的方差提升效果求平均。

场景 2 下筛选效果不是很明显,这可能是因为原始样本在筛选中占有更重要的地位。一旦固定原始样本的位置,那么原始样本附近的决策边界布局将确定,最终的优化难度也将基本确定。

越到最后的优化过程越困难,总体的优化难度更多取决于最终过程的优化难度。这也印证了一个普遍的实验现象,即对抗攻击的速度是先快后慢。也就是说,如果在分类 A 中选定了原始图片,与分类 B 中的众多目标图片形成的攻击组具有相似的攻击距离。

无论是哪种场景下,攻击组筛选后的攻击效果都是优于筛选前的攻击效果。筛选后攻击效率提高了 24.99%,方差降低了 76.23%,总体攻击效果提升了 42.07%,充分验证了基于决策边界长度的攻击距离度量方法的有效性,验证了攻击组筛选的有效性。

本文所提出的算法 1 与算法 2 是以图像识别领域为例,但均不局限于图像识别领域。这是因为:算法 1 的适用前提有 2 个:①分类器模型的输入样本可以测量样本间距离;②可以在样本附近进行高斯采样。如果这 2 个前提条件能够满足,那么算法 1 很容易复现,如语音识别领域的语音向量、恶意代码识别的代码向量^[33]等。算法 2 基于算法 1,没有任何额外的限制,因此本文提出的方法是适用于所有领域的分类器。同时,本文所提出的方法适用于所有对抗样本生成方法的攻击组选取。这是由于算法 1 与算法 2 的设计不局限于任何对抗样本生成方法,并在对比实验中证实了所提方法对多种对抗样本生成方法有效。

表 1 筛选前后查询次数对比

Table 1 Comparison of query times before and after filtering

场景	指标	NES Attack	Filtered NES	CMA Attack	Filtered CMA	平均提升效果/%
场景 1	平均值	1.01×10^5	7.87×10^4	5.81×10^4	2.88×10^4	36.25
	中位数	9.64×10^4	7.22×10^4	4.74×10^4	2.74×10^4	33.65
	方差	1.51×10^9	1.16×10^9	1.32×10^9	9.36×10^7	58.04
场景 2	平均值	9.79×10^4	9.66×10^4	4.75×10^4	4.26×10^4	5.82
	中位数	9.63×10^4	9.66×10^4	4.33×10^4	4.26×10^4	0.65
	方差	1.00×10^9	4.20×10^8	2.99×10^8	1.14×10^7	77.09
场景 3	平均值	9.76×10^4	6.64×10^4	5.92×10^4	3.05×10^4	40.22
	中位数	8.98×10^4	6.64×10^4	5.14×10^4	3.05×10^4	33.36
	方差	9.52×10^8	3.59×10^6	6.22×10^8	7.79×10^7	93.55

4 结 论

1) 针对攻击组间攻击难度不同导致攻击不稳定的问题,本文以图像识别领域为例,设计了基于决策边界长度的攻击距离度量方法,为攻击组的攻击难度提供了度量方法。

2) 设计了基于攻击距离的对抗样本攻击组筛选方法,筛去难以攻击的攻击组。通过预处理,提升现有算法的攻击效率和攻击稳定性。

3) 实验表明,筛选后的攻击组相比于筛选前,攻击效率提升了 24.99%,方差降低了 76.23%,总体攻击效果提升了 42.07%。

4) 建议所有利用攻击组的对抗样本生成方法,先进行攻击组的筛选,特别是对原始样本进行筛选,再展开攻击,以稳定攻击的效果。

为了更加清晰筛选对于攻击组的意义,笔者将会继续研究原始样本对于整个优化过程的意义。

参考文献 (References)

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6) : 84-90.

[2] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 1-9.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10) [2020-09-21]. <https://arxiv.org/abs/1409.1556>.

[4] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[5] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2261-2269.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19) [2020-09-21]. <https://arxiv.org/abs/1312.6199>.

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. (2015-05-20) [2020-09-21]. <https://arxiv.org/abs/1412.6572>.

[8] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[EB/OL]. (2017-02-11) [2020-09-21]. <https://arxiv.org/abs/1611.01236>.

[9] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. (2019-09-04) [2020-09-21]. <https://arxiv.org/abs/1706.06083>.

[10] DONG Y P, LIAO F Z, PANG T, et al. Boosting adversarial attacks with momentum[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.

[11] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C] // 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.

[12] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C] // 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.

[13] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: A simple and accurate method to fool deep neural networks[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2574-2582.

[14] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. (2017-05-19) [2020-09-21]. <https://arxiv.org/abs/1602.02697>.

[15] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2725-2734.

[16] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[EB/OL]. (2020-04-26) [2020-09-21]. <https://arxiv.org/abs/1705.07204>.

[17] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[C] // International Conference on Machine Learning, 2018: 2137-2146.

[18] KUANG X H, LIU H Y, WANG Y, et al. A CMA-ES-based adversarial attack on black-box deep neural networks[J]. IEEE Access, 2019, 7: 172938-172947.

[19] DONG Y P, SU H, WU B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition[C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7706-7714.

[20] BRENDEN W, RAUBER J, BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[EB/OL]. (2018-02-16) [2020-09-21]. <https://arxiv.org/abs/1712.04248>.

[21] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553) : 436-444.

[22] HELMSTAEDTER M, BRIGGMAN K L, TURAGA S C, et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina[J]. Nature, 2013, 500(7461) : 168-174.

[23] XIONG H Y, ALIPANAHI B, LEE L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218) : 1254806.

[24] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6) : 82-97.

[25] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C] // Annual Conference on Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2014: 3104-3112.

[26] MNH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540) : 529-533.

[27] GIUSTI A, GUZZI J, CIREŞAN D C, et al. A machine learning approach to visual perception of forest trails for mobile robots[J]. IEEE Robotics and Automation Letters, 2016, 1(2) : 661-667.

[28] MIDDLEHURST C. China unveils world's first facial recognition atm[EB/OL]. (2015-01-01) [2020-09-01]. <https://www.telegraph.co.uk/news/worldnews/asia/china/11643314/China-unveils-worlds-first-facial-recognition-ATM.html>.

[29] 刘恒, 吴德鑫, 徐剑. 基于生成式对抗网络的通用性对抗扰动生成方法[J]. 信息安全, 2020, 20(5) : 57-64.

LIU H, WU D X, XU J. Generating universal adversarial perturbations with generative adversarial networks[J]. Netinfo Security, 2020, 20(5) : 57-64 (in Chinese).

[30] HUANG Q, KATSMAN I, GU Z Q, et al. Enhancing adversarial example transferability with an intermediate level attack[C] // 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 4732-4741.

[31] WIERSTRA D,SCHAUL T,PETERS J,et al. Natural evolution strategies[C]//2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). Piscataway:IEEE Press,2008:3381-3387.

[32] HANSEN N. The CMA evolution strategy: A tutorial[EB/OL]. (2016-04-04) [2020-09-21]. <https://arxiv.org/abs/1604.00772>.

[33] 侯留洋,罗森林,潘丽敏,等. 融合多特征的 Android 恶意软件检测方法[J]. 信息安全学报,2020,20(1):67-74.

HOU L Y, LUO S L, PAN L M, et al. Multi-feature Android malware detection method[J]. Netinfo Security,2020,20(1): 67-74 (in Chinese).

A method for filtering the attack pairs of adversarial examples based on attack distance

LIU Hongyi, FANG Yutong, WEN Weiping*

(School of Software & Microelectronics, Peking University, Beijing 102600, China)

Abstract: During the generation of black-box adversarial examples, an attack pair is usually specified, including a source example and a target example. The purpose is to let the generated adversarial example only have little norm difference from the source example, but it is recognized by the classifier as the classification of the target sample. In order to solve the problem of the instability of adversarial attacks caused by different attack difficulty of attack pairs, taking the image recognition field as an example, this paper presented an attack distance measurement method based on the length of the decision boundary, which provided a measurement method for the attack difficulty of attack pairs. Then this paper designed a filtering method based on attack distance of the attack pairs, which filtered out attack pairs that were difficult to attack before the attack started, so this method can improve the attack performance without modifying the attack algorithm. Experiments show that, compared with the attack pairs before filtering, the filtered attack pairs improve the overall attack performance by 42.07%, improve the attack efficiency by 24.99%, and stabilize the variance by 76.23%. It is recommended that all methods of generating adversarial examples using attack pairs should filter attack pairs before attack to stabilize and improve the attack performance.

Keywords: adversarial examples; black box; decision boundary; filtering; image recognition

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0568

基于抽象汇编指令的恶意软件家族分类方法

李玉, 罗森林, 郝靖伟, 潘丽敏*

(北京理工大学 信息与电子学院, 北京 100081)

摘 要: 恶意软件变体的大量出现对网络安全造成巨大威胁。针对基于汇编指令的恶意软件家族分类方法中,操作数语义与运行环境密切相关而难以提取,导致指令语义缺失,难以正确分类恶意软件变体的问题。提出了一种基于抽象汇编指令的恶意软件家族分类方法。通过抽象出操作数类型重构指令,使操作数语义脱离运行环境的约束;利用词注意力机制与双向门循环单元(Bi-GRU)构建指令嵌入网络以捕获指令行为语义,并结合双向循环神经网络(Bi-RNN)学习恶意软件家族共性指令序列,以减小变体技术对指令序列的干扰;融合原始指令和家族共性指令序列构建特征图像,并通过卷积神经网络实现恶意软件家族分类。公开数据集上的实验结果表明:所提方法能够有效提取操作数信息,抵抗恶意软件变体中无关指令的干扰,实现恶意软件变体的家族分类。

关 键 词: 恶意软件家族分类;可视化;抽象汇编指令;卷积神经网络;双向循环神经网络(Bi-RNN);词注意力机制

中图分类号: TP309.5

文献标志码: A **文章编号:** 1001-5965(2022)02-0348-08

恶意软件是指在未经用户许可的情况下收集敏感信息、控制用户设备、严重侵犯用户个人权益的软件^[1]。目前,恶意软件检测方法根据检测时是否执行目标代码,分为静态分析^[2-3]和动态分析^[4-5]。随着深度学习在图像识别领域展现出强大的特征提取能力,并且归属于同一家族的恶意软件变体,其特征图像的纹理特征也十分相似,因此可视化技术开始应用在恶意软件任务中。根据可视化的目标,恶意软件可视化检测方法分为二进制文件可视化和动态行为可视化。二进制文件可视化将文件按字节排序,生成特征图像并识别所属家族^[6-8],并未提取深层行为特征,仍依赖于待检测软件的二进制代码,易受到代码混淆的影响。

为了将恶意行为信息融入特征图像中,汇编指令逐渐成为可视化研究的重心。鉴于二进制文

件存在加壳现象,难以自主生成反汇编文件,因此对于加壳文件需通过动态调试进行脱壳,再进行反汇编。获得汇编代码后,Zhang等^[9]提出了基于图像识别恶意的软件检测方法(IRMD),从反汇编文件中提取操作码序列,构建2-gram操作码特征组,采用卷积神经网络(CNN)进行识别和分类,证明了操作码序列具有较强的行为特征表达能力,且同一家族的恶意软件变体在汇编指令层面具有相似的行为模式。受IRMD启发,Ni等^[10]开发了基于局部敏感哈希的恶意软件分类方法(MSCS)。MSCS将每个操作码视为一个序列,以关注软件的整体行为特征,并引入局部敏感哈希算法生成特征图像,利用CNN进行软件的恶意性检测。与MSCS类似,Sun和Qian^[11]通过循环神经网络(RNN)捕获指令序列内的时序特征,并对

收稿日期: 2020-09-30; 录用日期: 2020-11-13; 网络出版时间: 2021-02-14 15:59
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210210.1745.002.html
基金项目: 工信部2020年信息安全软件项目(CEIEC-2020-ZM02-0134)
* 通信作者. E-mail: panlimin2016@gmail.com

引用格式: 李玉, 罗森林, 郝靖伟, 等. 基于抽象汇编指令的恶意软件家族分类方法[J]. 北京航空航天大学学报, 2022, 48(2): 348-355. LI Y, LUO S L, HAO J W, et al. Malware family classification method based on abstract assembly instructions[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 348-355 (in Chinese).

指令序列进行预测,一定程度上减小了变体技术对指令序列的影响。

上述可视化检测方法将软件行为融入特征图像中,但仅从操作码层面提取行为特征,并未结合操作数,导致指令语义缺失。同时,恶意软件变体通常在软件中插入无关的正常指令以隐藏恶意指令和恶意行为,且正常指令与恶意指令仅通过操作码难以区分,导致基于操作码序列的检测方法难以区分出无关指令,受代码混淆和花指令干扰较大,影响变体分类结果。

经分析,无关指令和恶意指令的区别仅在操作数上有所体现,但操作数的语义与运行环境密切相关,即同一个操作数在不同运行环境下的含义不同,因此难以提取。

针对基于汇编指令的恶意软件可视化家族分类方法中,操作数语义与运行环境密切相关而难以提取,导致指令语义缺失,难以正确分类恶意软件变体的问题,本文提出了一种融合抽象汇编指令和家族共性指令序列的恶意软件家族分类方法 (malware family classification method based on ab-

stract assembly instructions, MCAI), 利用操作数抽象类型提取进行指令重构,使操作数语义脱离运行环境的约束而易于提取,并结合词注意力机制和双向门循环单元 (Bi-GRU) 构建双向结构指令嵌入网络,学习指令的行为语义,并利用深度神经网络捕获恶意软件家族共性指令序列,以减小恶意软件变体技术对指令序列的干扰,融合原始指令和家族共性指令序列构建特征图像,并通过 CNN 实现恶意软件家族分类。

1 MCAI 原理

1.1 MCAI 原理框架

MCAI 原理如图 1 所示,包含指令重构模块、指令嵌入模块、家族共性指令序列提取模块、特征图像生成模块和特征图像分类模块。指令重构模块通过操作数抽象类型提取重构指令,使操作数语义脱离运行环境的约束,丰富指令的行为语义,并对重构后的指令进行词法分析,识别指令内部的各个元素,以便后续特征提取;指令嵌入模块利用词注意力机制和 Bi-GRU 构建指令嵌入网络,

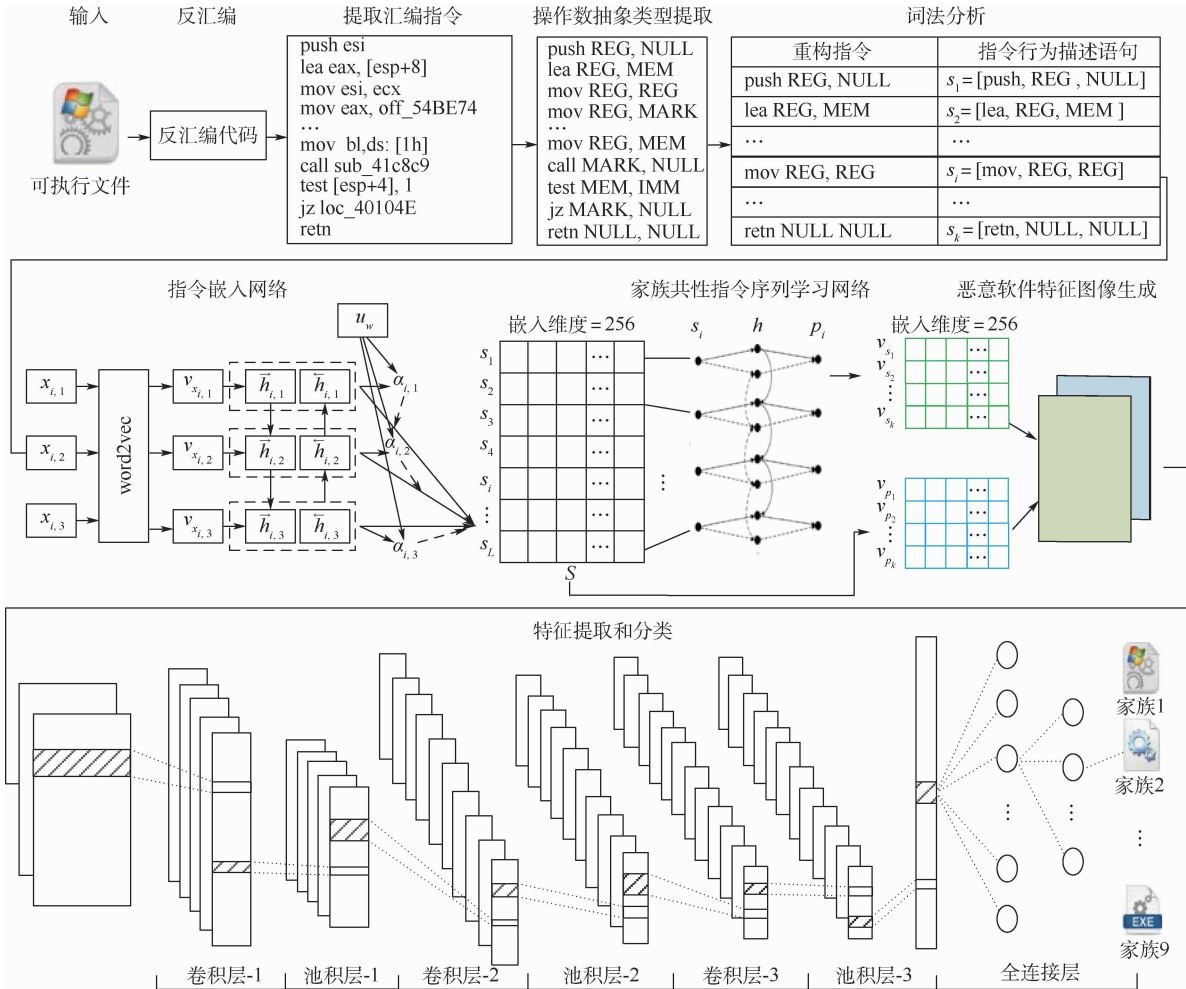


图 1 MCAI 原理框图

Fig. 1 Functional block diagram of MCAI

学习指令的行为语义,获得指令的向量表示,并通过深度层次网络学习家族共性行为模式,对指令进行预测;特征图像生成模块将预测结果和嵌入结果组合成双通道恶意软件特征图像;特征图像分类模块利用 CNN 捕获图像的局部特征,并给出家族分类结果。

1.2 指令重构模块

汇编指令是程序在操作系统层面的基本执行单元,因此能够反映程序的操作和意图。通过反汇编器,可以获得样本的反汇编代码 p 。其中, p 由多条指令 I_i 组成, $i \in [1, k]$ 。若恶意软件已被加壳,需先经过脱壳处理,再进行上述反汇编过程。

x86 指令语法规则规定,汇编指令 I_i 由一个操作码和多个操作数组成。操作码表示指令的行为类型,操作数体现了指令的操作目标。因此,汇编指令中的每一个组成元素都是指令行为的一部分,仅通过操作码或操作数无法完整描述指令行为。然而,操作数的语义存在不确定性,同一个地址在不同的软件中可能表示不同的函数,因此在过去的检测和分类方法中,如基于 RNN 的恶意软件可视化分类方法 (RMVC)、MCSC 等,仅保留了操作码序列作为特征,导致无法区分出恶意软件变体内的正常指令,影响分类效果。为了充分利用操作数信息,使操作数的语义脱离运行环境的限制,方法对操作数类型进行抽象,定义了内存地址 (MEM)、寄存器 (REG)、立即数 (IMM)、标号 (MARK, 包括函数名、跳转地址等) 和无操作数占

位符 (NULL) 5 类操作数类型,用操作数类型代替操作数完成指令重构。

由于每条汇编指令内操作码数量不一致,对操作数数量进行规范化处理,即通过 NULL 实现指令对齐,使每条指令的操作数数量均为 2 个,补齐后的指令集如图 2 所示。

对规范化的指令 I'_i 进行词法分析,识别指令内部的各个元素,并构造指令行为描述序列 s_i 。其中, s_i 由 3 部分组成,分别为操作码和 2 个抽象操作数类型。

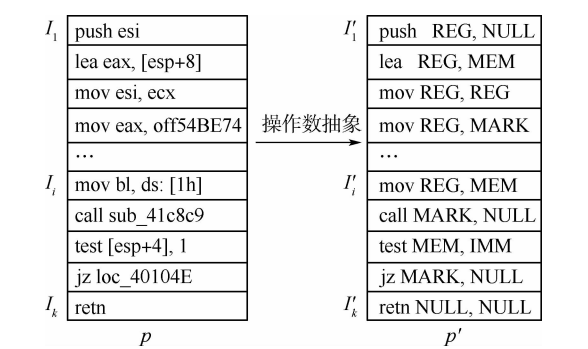


图 2 操作数类型抽象
Fig. 2 Operand type abstracting

1.3 指令嵌入模块

指令嵌入网络将每个指令行为描述序列嵌入至其向量表示中,并学习指令的行为语义,以便特征提取网络可以提取行为特征。指令嵌入网络由 word2vec^[12]、Bi-GRU^[13] 和词注意力模块^[14] 构建,如图 3 所示。

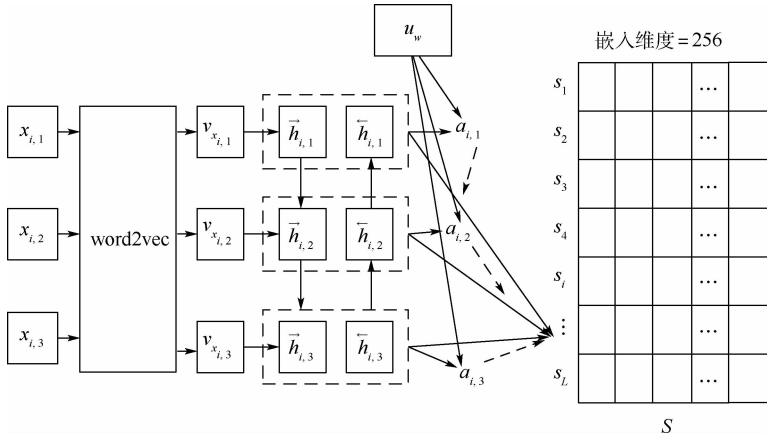


图 3 指令嵌入网络
Fig. 3 Instruction embedding network

利用 word2vec 构建词嵌入网络,训练 s_i 中每一个单词 $x_{i,j}$ 的向量表示。

在词向量基础上,利用 Bi-GRU 构建句子嵌入网络以训练每条指令的向量表示。鉴于指令内部的序列信息为双向关联,因此采用 Bi-GRU 分别从前向和后向捕获指令内部的序列信息,并通

过词注意力机制为指令内的各个元素自动化分配权重,以进一步突出指令的关键信息,增强指令的行为语义。

指令嵌入网络从 3 个方面学习指令的行为语义信息。word2vec 提取每个单词的基本上下文信息, Bi-GRU 分别从前向和后向学习了输入指令内

部各元素的顺序信息,词注意力模块从指令内元素的上下文中学习元素的权重,并加权得到指令最终的嵌入结果。

1.4 家族共性指令序列提取模块

属于同一家族的恶意软件具有相同的行为特征,在汇编代码中体现为具有相似的指令序列,并且指令序列中具有较强的时序性。双向循环神经网络 (Bi-RNN)^[15] 具有较强的时序特征提取能力,通过 Bi-RNN 可以学习恶意软件家族的通用行为特征。

为了更好地说明 Bi-RNN 的学习过程,图 4 中用标号代替指令,每一个标号对应一条指令。

Bi-RNN 采用滑动窗口的方式学习指令的时序特征。在 t 时刻,Bi-RNN 学习虚线框内的指令序列并预测圆圈内的指令内容,得到 p_t 。在 $t+1$ 时刻,滑窗右移一个指令,再次学习窗口内的指令序列并预测圆圈的指令内容,得到 p_{t+1} 。随着滑窗的移动,指令的预测值组成一个新的指令序列,即为原指令序列的预测结果,损失函数即为指令的预测值和指令的真实值之间的差异。在反向传播阶段,为了快速降低损失函数值,滑动窗口的预测结果将会是窗口中最常见的指令,从而使 Bi-RNN 学习到该家族通用的指令序列。因此,即使恶意软件通过变体技术调整的部分指令,Bi-RNN 也可以根据指令序列的上下文信息还原指令序列,有效减小了恶意软件变体技术对指令序列的影响,如图 5 所示。

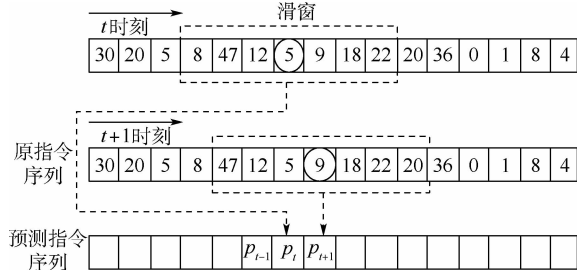


图 4 家族共性指令序列学习

Fig. 4 Family common instruction sequence learning

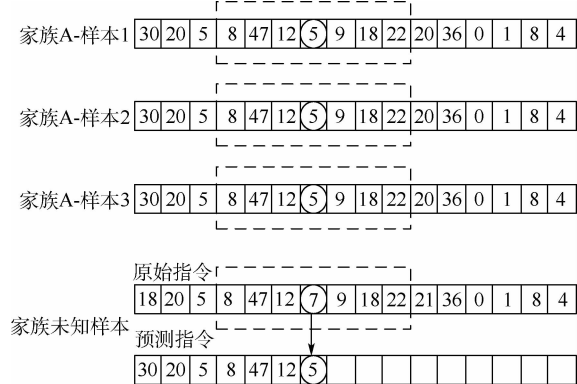


图 5 家族共性指令序列预测

Fig. 5 Family common instruction sequence prediction

预测指令通过查询嵌入表可得到嵌入结果 v_{p_t} 。

1.5 特征图像生成模块

为充分利用 CNN 的特征提取能力,将 Bi-RNN 的预测结果分别转换为单通道图像,并组成双通道的恶意软件特征图像。特征图像生成过程如下:

- 1) 将嵌入结果和预测结果的数值归一化至 $0 \sim 255$,对应灰度图像像素值。
- 2) 顺序排列,分别生成单通道图像 $M \in \mathbf{R}^{1 \times k \times w}$ 。其中, k 为样本内指令个数, w 为指令嵌入的输出维度,实验中的嵌入维度为 256。
- 3) 鉴于样本内的指令数量不同,因此生成的图像长度不一致,需要对 2 个单通道图像各自放缩。通过双线性差值法,将 2 个单通道图像均放缩大小为 $3\,000 \times 256$ 。其中,256 为指令嵌入的输出维度,3 000 为数据集内指令数量的中位数。
- 4) 2 个放缩后的单通道图像组成双通道恶意软件特征图像 $M \in \mathbf{R}^{2 \times 3\,000 \times 256}$ 。

1.6 特征图像分类模块

通过 CNN 可以捕获图像内的局部特征。鉴于图像中的每一行对应一个指令,为确保一条指令的信息可以被完整的处理,在第一个卷积层中,使用宽卷积核 $f \in \mathbf{R}^{2 \times h \times 256}$,其中 h 为卷积核高度,代表一次处理指令的个数,实验中取 $h = 2$,即经过第一个卷积层后,每个通道的卷积结果为列向量。随后,该卷积结果依次经过其他卷积层和池化层,完成图像的局部特征提取,并由全连接层和 softmax 层进行分类。

2 实验分析

2.1 实验数据

实验数据来源于 2015 Kaggle 微软恶意软件分类挑战赛^[16],数据集中包含来自 9 个家族的 10 868 个有标记样本,包含样本的二进制文件和反汇编文件,是目前恶意软件研究领域广泛使用的数据集。数据集家族标签及样本数量如表 1 所示。

表 1 数据集家族标签及样本数量

Table 1 Family label and number of the sample in the dataset

序号	家族名称	样本数量
1	Ramnit	1 541
2	Lollipop	2 478
3	Kelihos_ver3	2 942
4	Vundo	475
5	Simda	42
6	Tracur	751
7	Kelihos_ver1	398
8	Obfuscator. ACY	1 228
9	Gatak	1 013

示。鉴于本文方法是基于汇编代码实现的,因此仅使用反汇编文件。训练集占 40%,验证集占 10%,测试集占 50%。

2.2 实验环境

实验硬件环境为: Intel (R) Core (TM) i7-6700 CPU@ 3. 40 GHz, RAM 8 GB, Windows10 系统。主要软件和开发工具包为: Python 3. 7、TensorFlow v1. 14. 0、scikit-learn0. 22。

2.3 评价方法

实验采用机器学习领域的规范统计指标:准确率、召回率、精确率及 F_1 ,来准确评估分类模型的性能,从而更加全面地体现方法的有效性和可靠性。

2.4 消融实验

2.4.1 消融实验目的

本节实验通过拆分实验过程,以确认方法中每一个环节的有效性和意义,具体包含如下 3 点:

- 1) 验证操作数类型抽象和指令重构意义。
- 2) 验证指令嵌入网络对指令行为语义捕获的贡献。
- 3) 验证家族共性指令序列对恶意软件变体分类能力的提升。

2.4.2 消融实验过程

为验证指令重构、指令嵌入和学习通用指令行为序列的效果,对比了 5 种恶意软件家族分类方法:

- 1) 以操作码序列作为样本特征,采用 256 位哈希编码实现操作码向量化并排列成 16×16 恶意软件特征图像,通过 CNN 识别恶意软件家族。
- 2) 以操作码和操作数组成的原始指令作为样本特征,采用 256 位哈希编码实现指令向量化并排列成 16×16 图像,通过 CNN 识别恶意软件家族。
- 3) 以操作码和抽象操作数类型组成的重构指令作为样本特征,采用 256 位哈希编码实现指令向量化并排列成 16×16 图像,通过 CNN 识别恶意软件家族。
- 4) 以操作码和抽象操作数类型组成的重构指令作为样本特征,采用指令嵌入网络实现指令向量化并排列成图像,图像大小为 $3\,000 \times 256$,通过 CNN 识别恶意软件家族。
- 5) 以操作码和抽象操作数类型组成的重构指令作为样本特征,采用指令嵌入网络实现指令向量化,通过 Bi-RNN 学习家族共性指令序列并做出指令预测,将重构指令与预测指令合并为双通道特征图像,图像大小为 $2 \times 3\,000 \times 256$,通过 CNN 识别恶意软件家族,即 MCAI。

其中,方法 1 ~ 方法 3 网络结构相关参数见 2.5.2 节,方法 4、方法 5 中 CNN 卷积层采用 2×2 卷积核,池化层采用 2×2 池化窗口,其余结构见 2.5.2 节。

2.4.3 消融实验结果

消融 5 种恶意软件家族分类方法 1 ~ 方法 5 实验结果如表 2 所示。

表 2 消融实验结果				
Table 2 Results of ablation experiments				
方法	准确率	召回率	精确率	F_1
1	91.49	88.87	87.81	88.32
2	80.99	71.53	69.53	70.28
3	92.49	88.04	87.27	87.57
4	96.04	91.13	96.28	93.01
5	98.51	98.36	97.07	97.66

实验结果表明,消融方法 1 ~ 方法 5 在准确率方面分别达到了 91.49%、80.99%、92.49%、96.04% 和 98.51%。方法 1 仅使用操作码序列作为样本特征,准确率达到 91.49%,表明操作码序列可以在很大程度上反映恶意软件行为。但是在方法 2 中,结合了原始操作数后,实验效果显著下降,表明操作数的含义是每个恶意软件样本的特有信息,结合过多的特有信息将导致过拟合。方法 3 以操作数类型替换原始操作数实现指令重构,准确率有了较为明显的提升,表明操作数类型有助于指令行为的描述和提取。方法 4 采用指令嵌入网络实现指令向量化,相比于方法 3 中的哈希编码有较大的提升。经分析,哈希编码仅关注指令的整体信息,无法提取指令内部各元素之间的关系,而指令嵌入网络分别从词、句 2 个层面提取指令内、外的上下文信息,因此可以更好地学习指令特征,保留指令的行为语义。方法 5 通过 Bi-RNN 学习指令通用行为模式,对每条指令进行预测,并将预测结果和原始指令合并为双通道特征图像,实现效果优于方法 1 ~ 方法 4,表明 Bi-RNN 能够有效学习到指令的序列信息,捕获指令通用的序列模式。通过合并预测结果和原始指令,加强了属于同一家族的变体的相似性,可以在一定程度上减小变体技术的影响。

2.5 对比实验

2.5.1 对比实验目的

为了验证 MCAI 在恶意软件家族分类任务上的有效性,将 MCAI 与可视化检测领域内的优秀算法 RMVC^[11]、MCSC^[10] 进行比较。

2.5.2 对比实验过程

实验过程如图 6 所示。实验中,设置 batch size 为 128,学习率为 0.000 8,设置迭代次数为 10 000,

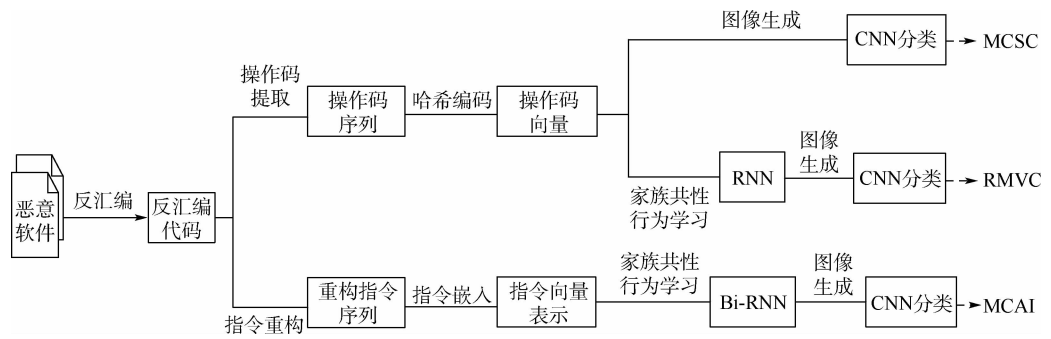


图6 对比实验流程

Fig.6 Flowchart of comparative experiments

L_2 正则化系数为 0.01,词嵌入维度为 256。CNN 中,卷积层 1 采用 32 个 3×256 大小卷积核,卷积层 2、3 均采用 64 个 3×1 大小卷积核,所有池化层均采用 4×1 大小池化窗口,全连接层神经元个数分别为 512、256、9。

2.5.3 对比实验结果

实验结果如表 3 所示。图 7 ~ 图 9 显示了 3 种方法的精确率、召回率和 F_1 值。可以看到,

表3 对比实验结果

方法	准确率	召回率	精确率	F_1
MCSC	91.4	88.87	87.81	88.32
RMVC	94.20	95.92	89.07	91.00
MCAI	98.51	98.36	97.07	97.66

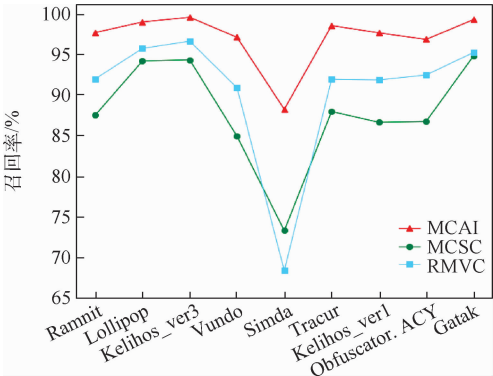


图9 对比实验召回率

Fig.9 Recall of comparison experiment

本文方法在各项指标上均有所提升。此外,数据集集中的 Simda 家族有 42 个样本,占样本总数的 0.39%,处于严重不平衡状态。从实验结果中看出, RMVC 在不平衡数据上的效果较差,受数据分布的影响较大; MCSC 则表现较好,能一定程度应对数据不平衡问题;而本文方法在 Simda 家族上的分类结果达到了 96%,对数据不平衡问题有较强的鲁棒性。

3 结 论

针对基于汇编指令的恶意软件家族分类方法操作数语义与运行环境密切相关而难以提取,导致指令语义缺失,难以正确分类恶意软件变体的问题,本文提出了一种融合抽象汇编指令和家族共性指令序列的恶意软件家族分类方法。

1) 通过操作数类型抽象重构汇编指令,使操作数语义脱离运行环境的约束,解决了操作数语义依赖于运行环境而难以提取的问题,进一步增强了指令行为描述的充分性和完整性。

2) 通过 word2vec、Bi-GRU 和词注意力机制构建指令嵌入网络,实现指令向量化,并分别从指令内和指令间学习上下文信息,使指令的向量表示中包含更丰富的行为语义。

3) 通过 Bi-RNN 学习指令的序列信息,捕获

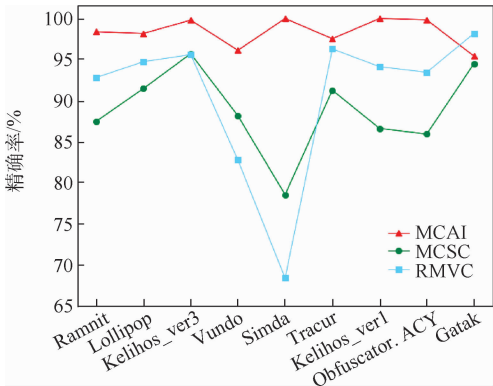


图7 对比实验精确率

Fig.7 Precision of comparison experiment

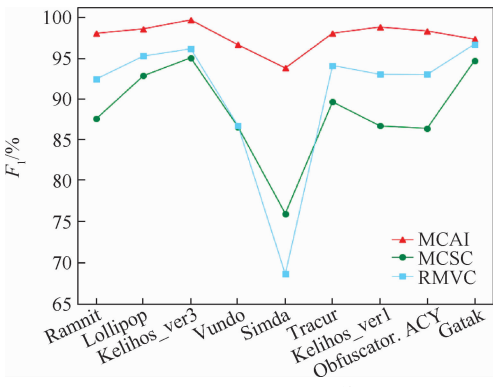


图8 对比实验 F_1 值

Fig.8 F_1 -score of comparison experiment

家族共性指令序列,并对每条指令给出预测指令,能够一定程度上发现并还原变体技术对指令的干扰和微调,减小代码混淆等变体技术对指令序列的影响。同时,结合预测指令和原始指令构建双通道恶意软件特征图,加强了属于同一家族的变体的相似性,进一步提升了家族分类效果。

参考文献 (References)

[1] YE Y F, LI T, ADJEROH D, et al. A survey on malware detection using data mining techniques[J]. ACM Computing Surveys, 2017, 50(3) : 1-40.

[2] YE Y F, LI T, ZHU S H, et al. Combining file content and file relations for cloud based malware detection[C] // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York : ACM Press, 2011 : 222-230.

[3] TAMERSOY A, ROUNDY K, CHAU D H. Guilt by association: Large scale malware detection by mining file-relation graphs[C] // Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York : ACM, 2014 : 1524-1533.

[4] DING Y X, XIA X L, CHEN S, et al. A malware detection method based on family behavior graph[J]. Computers & Security, 2018, 73 : 73-86.

[5] HARDY W, CHEN L, HOU S, et al. D14md: A deep learning framework for intelligent malware detection[C] // Proceedings of the International Conference on Data Mining(DMIN). New York : CSREA, 2016 : 61.

[6] NATARAJ L, KARTHIKEYAN S, JACOB G, et al. Malware images: Visualization and automatic classification[C] // Proceedings of the 8th International Symposium on Visualization for Cyber Security. New York : ACM, 2011 : 1-7.

[7] CUI Z H, XUE F, CAI X J, et al. Detection of malicious code variants based on deep learning[J]. IEEE Transactions on Industrial Informatics, 2018, 14(7) : 3187-3196.

[8] TRINIUS P, HOLZ T, GÖBEL J, et al. Visual analysis of malware behavior using treemaps and thread graphs[C] // 2009 6th International Workshop on Visualization for Cyber Security. Piscataway : IEEE Press, 2009 : 33-38.

[9] ZHANG J X, QIN Z, YIN H, et al. IRMD: Malware variant detection using opcode image recognition[C] // 2016 IEEE 22nd International Conference on Parallel and Distributed Systems(ICPADS). Piscataway : IEEE Press, 2016 : 1175-1180.

[10] NI S, QIAN Q, ZHANG R. Malware identification using visualization images and deep learning[J]. Computers & Security, 2018, 77 : 871-885.

[11] SUN G S, QIAN Q. Deep learning and visualization for identifying malware families[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(1) : 283-295.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[C] // Proceedings of ICLR Workshops Track. Cambridge : MIT Press, 2013 : 21-29.

[13] CHO K, VAN MERRIENBOER B, BAHDANAU D, et al. On the properties of neural machine translation: Encoder-decoder approaches[C] // Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014 : 103-111.

[14] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016 : 33-38.

[15] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11) : 2673-2681.

[16] Microsoft. Microsoft malware classification challenge (Big 2015) [EB/OL]. (2019-06-11) [2020-09-01]. <https://www.kaggle.com/c/malware-classification>.

Malware family classification method based on abstract assembly instructions

LI Yu, LUO Senlin, HAO Jingwei, PAN Limin*

(School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: The emergence of malware variants poses a great threat to network security. In malware family classification methods based on assembly instructions, the semantics of operands are closely related to the operating environment and difficult to extract, which leads to the lack of instruction semantics and the difficulty in correctly classifying malware variants. A malware family classification method based on abstract assembly instructions is proposed. The instruction is reconstructed by abstracting the operand type, so that the semantics of the operands can be separated from the constraints of the operating environment. The word attention mechanism and bidirectional gate recurrent unit (Bi-GRU) are used to construct an instruction embedding network and to capture the instruction behavior semantics. Combined with bidirectional recursive neural networks (Bi-RNN), the common instruction sequence of malware family is learned to reduce the interference of variation technology on the instruction sequence. The original instruction and family common instruction sequence are integrated to construct feature images, and the malware family classification is realized through convolutional neural network. The experimental results on the public dataset show that the proposed method can effectively extract operand information, resist the interference of irrelevant instructions in malware variants, and realize the family classification of malware variants.

Keywords: malware family classification; visualization; abstract assembly instructions; convolutional neural network; bidirectional recursive neural network (Bi-RNN); word attention mechanism

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0552

基于关键字的海报自动合成系统

关帅鹏, 于海阳, 杨震*, 周明, 赖英旭

(北京工业大学 信息学部, 北京 100124)

摘 要: 智能化的普及对图像编辑提出了新需求,海报作为一种以图像形式传递信息的方式,在日常生活和工作管理中起着重要的作用。海报的制作需要多元素图像进行合成,目前缺少一种交互式的、一键式的图像合成系统,因此,结合当前流行的图像处理技术,设计并实现了一款海报自动合成系统。提出了一种基于关键字的图像检索方案,构建基于文本和内容的双重过滤方案,为用户提供精准快捷的图像检索手段;通过对大量精心设计的海报图像统计构图规律并引入美学常识的构图规则,提出了一种基于双向规则的人像布局推荐方案;在双向规则的共同作用下辅助用户进行人像布局设计。实验结果表明:所提方案能够稳定高效地运行,用户能够通过简单的交互操作实现图像合成,最终图像合成的效果真实有效。

关 键 词: 图像检索; 自动分割; 无缝融合; 布局推荐; 自动合成
中图分类号: V221+.3; TB553
文献标志码: A **文章编号:** 1001-5965(2022)02-0356-13

海报被视为一种强有力的宣传载体,在大众日常生活、生产管理及科学研究等领域扮演着重要的角色。主题明确、精心设计的海报不但能够吸引观众的注意力,而且对于主题信息的表达,也能够准确完整的呈现。海报设计中,使用图像编辑工具辅助设计,这与图像合成的理念十分契合,即从一张或多张图像中分割场景项并将其无缝地粘贴到另一张背景图像内,通过对图像、文字、空间等要素进行完整的结合,使得背景图像获得额外的信息传递效果。与普通图像相比,合成的图像能在时间和空间上获得额外的表述,能够更加完整地表达图像需求者的意愿。Johnson 等^[1]实现了一种图像合成系统,通过定义文字标注和文本元素区域的方式,把标注后的图像存入数据库,在数据库内检索相关图像并根据画布提供的位置信息合成目标图像,该方法从标注的数据库中检

索图像,仅适用于风景类的图像合成,适用范围小,且用户无法自适应地改变图像合成结果。Chen 等^[2]使用手绘草图和文字标注的组合信息从互联网搜索引擎中检索图像,并将图像元素按照手绘草图中的布局信息拼接至背景的对应位置实现图像合成,该方法忽略了前景图像间的布局信息,最终合成结果中不可避免地会出现肉眼可见的逻辑错误。

海报图像合成主要涉及 2 个问题,即如何准确地获取目标图像,以及如何分割将其无缝地糅合到目标图像中。根据图像检索中启发性信息不同,可将图像检索大致分为基于文本和基于内容的检索方法。基于文本的检索方法实质上把图像检索映射成文本检索的方式,对图像内容反映的信息理解进行文本标注,实现图像与文本的信息等价,通过文本检索实现图像检索。基于内容

收稿日期: 2020-09-27; 录用日期: 2020-12-04; 网络出版时间: 2021-03-18 10:17
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20210317.1118.001.html
基金项目: 国家自然科学基金(61671030); 北京市长城学者计划(CIT&TCD20190308); 中国博士后科学基金(2019M660377)
* 通信作者: E-mail: yangzhen@bjut.edu.cn

引用格式: 关帅鹏, 于海阳, 杨震, 等. 基于关键字的海报自动合成系统[J]. 北京航空航天大学学报, 2022, 48(2): 356-368.
GUAN S P, YU H Y, YANG Z, et al. Automatic poster synthesis system based on keywords[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 356-368 (in Chinese).

的图像检索方法不再使用文本信息作为描述符,而是结合特定算法提取图像颜色、纹理、轮廓等一系列内容特征,根据图像本身的内容进行检索,这种对图像检索的处理技术更准确。世界上第一个基于内容的商用图像检索系统是 IBM 的 QBIC 系统^[3],该系统的语义识别是通过人工完成的,导致许多图像的度量化比较差。麻省理工学院也在基于内容图像检索领域研发了图像检索系统,即 Photobook^[4]。针对从源图像中分割场景项并将其无缝地粘贴到目标图像的问题,Agarwala 等^[5]提出了一个经典的、交互式的计算机辅助图像合成框架,框架包括 2 部分:使用图像分割(image segmentation)提取候选图形,紧接着使用图像融合实现无缝融合(seamless cloning)。在图像检索完之后,如何对图像进行分割处理也是重要的研究热点之一。Bezdek^[6]在图像分割中使用了聚类的方法,通过图像像素间的相似程度把图像划分成若干区域子集,该方法的缺点在于大多数算法需要事先定义簇的个数(图像分割块数),并且特征选择单一,没有考虑像素的细节结构,分割效果不尽人意。Boykov 和 Jolly^[7]利用图分割模型对能量函数进行最优化求解,提出了一种交互式图像分割算法,通过用户手动指定像素所属类,来实现图像的二值分割。在此基础上,Rother 等^[8]提出了 GrabCut 算法,该算法引入高斯混合模型对前景、背景元素进行建模,进一步简化了使用者的操作难度,降低了对图像分割的计算量。Felzenszwalb 和 Huttenlocher^[9]在图论的基础上研究如何进行图像分割,提出了一种基于最小生成树的改进算法,将同一子图内部的差异性和子图之间的差异性相组合,实现像素在聚类时的自适应阈值选择,形成若干个彼此差异最大的临界子图,完成对图像的分割处理。图像无缝融合是指将 2 个或者 2 个以上的源图像进行糅合,达到对同一场景图像的完美表达。目前,比较流行的有基于 Alpha 和基于梯度场的无缝融合算法。Wang 和 Cohen^[10]利用掩码分割目标区域,并将目标图像区域与背景图像根据 α 分量(图像透明度)进行叠加获得无缝的融合结果。在前景和背景图像光照和纹理特征相差较大时,Alpha 融合算法将会产生明显的人工融合间隙。为了解决上述问题,Fattal 等^[11]提出了梯度域的图像编辑方法,通过求解图片梯度场和对比值的压缩,构建泊松方程的形式,实现在动态范围内图像的压缩。在此基础上,Pérez 等^[12]提出了泊松图像编辑方法,通过狄利克雷边界条件达到对泊松方程的求解。在保

留前景图像梯度信息的同时,保证图像在梯度域上的连续,这样就保证了前景和背景图像间光照和纹理特征一致。

除上述 2 个典型问题,本文在系统中额外加入人像布局推荐来辅助对以人像为主题的图像进行构图设计。人像布局推荐是一个新兴的研究领域,前景人像区域和背景区域之间的位置、大小等布局信息关系到合成图像是否真实合理。人像布局推荐的目标是通过调整前景图像、背景图像间相对的大小、位置信息将前景图像毫无违和感地嵌入另一张背景图像中,使图像合成更真实。在图像编辑中,设计图像的布局结构属于开放性问题,具有一定的难度。Inaba 等^[13]利用场景间关系确定元素项位置,通过计算数据库中 111 个场景标签间距离,统计出关键字标签的概率分布,达到生成元素位置的效果。Bhattacharya 等^[14]通过学习大量用户数据来获取图像美学的支持向量回归模型,通过空间重组来提高照片的视觉美感,用于辅助用户改进图像的构图选择。Zhang 等^[15]通过构建大量专业照片提取代表美学空间构图的人像注意力和几何构图特征。根据背景的相似性,为新的背景自适应地挑选相似背景下的人像作为参考图像,实现人像姿态和位置的推荐。Wang^[16]和 Rawat^[17]等在此基础上,收集上千幅构图良好的人像照片,从构图良好的图像中学习人像的审美构图规律,结合美学构图规则,同时利用背景内动态视觉元素实现版面平衡的方法,构建了一个辅助摄影系统来实现针对不同背景下的人像布局推荐。

1 系统描述

本文设计了一个基于关键字的海报自动合成系统,其整体框架如图 1 所示。按照合成流程可将系统分为以下 3 个阶段。

- 1) 图像检索阶段。主要是对目标图像的前景、背景图像的准备,用户通过提供关键字标签或文本信息来进行相关目标图像的提取,为后续图像融合提供素材。
- 2) 图像布局设计阶段。用于辅助用户对人像元素进行布局设计,优化图像构图,保证图像信息准确快速的表达。考虑使用基于统计的正规则和基于美学的负规则进行双向的线性组合来量化背景下的人像布局评分,实现人像布局推荐。
- 3) 图片融合阶段。是系统的最后一步,按照布局设计的方案将图像素材检索提供的前景和背景图像进行无缝融合即可实现图像合成。将无缝



图1 海报自动合成系统流程

Fig. 1 Procedure of automatic poster synthesis system

融合的结果返回给用户,以供用户选择。

系统需要建立 2 个数据库:第 1 个是通过互联网搜索引擎实时获取前景、背景图像的在线数据库,用于图像检索阶段;另 1 个是事先建立的离线海报数据库,用于图像布局设计阶段。根据上述对系统的分析,基于关键字的海报自动合成系统主要由以下模块组成:

1) 图像下载。图像下载模块负责对目标图像、背景元素等的获取。系统依托互联网搜索引擎,根据用户提供的文本信息下载对应的图像数据,按照特定规则将图像存储建立数据库,为后续的复杂性过滤和一致性排序提供数据源。

2) 复杂性过滤。复杂性过滤模块设置的目的是:一方面可以过滤错误图像,提升检索效率;另一方面还能消除内容复杂的图像,保证在后续图像分割和融合过程中产生更好的效果。

3) 一致性排序模块。一致性排序通过提取图像内容的多维度特征,并对图像特征进行内容一致性排序来达到二次检索的效果,进一步提升检索的性能。

4) 人像布局推荐模块。人像布局推荐可辅助用户进行海报的布局设计。从构图规律和美学常识的角度,为不同的人像与背景的组合生成恰当的布局备选。

5) 图像融合模块。图像融合是将从前景目标区域无缝融合至背景图像中的过程,提供最终的海报合成结果。

本文将自动的图像检索和人像布局推荐作为系统的重点核心内容,将围绕上述各模块进行深

入的分析和人性化的设计。

2 基于关键字的图像检索与融合

本文采用基于信息熵差的关键字提取算法^[18]进行文本处理。该算法提出文章关键词是以聚类的形式存在的,按照每个词出现间距是否大于阈值划分为内外 2 种模式,通过计算 2 种模式下的信息熵差来实现关键字提取。

关键词标签一旦确定,通过互联网搜索引擎下载对应图像。在后续的工作中,将按照特定的标准分别处理背景和前景图像。首先,是内容一致性,即不论是前景还是背景图像,图像内容都应该与对应的关键字标签一致;其次,考虑内容复杂性,对于背景图像,图像内容应该尽可能保持简单,为前景元素的加入保留足够的开放空间;而前景图像涉及后续目标区域的自动分割,应该尽量选择算法自动分析相对可靠的图像。在图像下载过程的同时,对收集到的所有图像进行预处理,删除下载失败或像素较低的模糊图像,为最终的图像合成提供高质量的图像素材。

2.1 复杂性过滤算法设计

在进行复杂性过滤算法设计时,本文使用图的分割算法^[9]来进行语义分割。该算法是一种基于图的贪心聚类算法,不仅考虑了同一子图内部的差异性,还考虑了子图与子图间的差异性。算法的关键在于像素聚类时能够实现自适应阈值。然后使用一种基于短连接的深度学习显著性检测算法^[19]。该算法通过在 HED (holistically-

nested edge detection) 架构内引入跳层结构的短连接,充分利用了从全卷积网络提取的多层次和多尺度特征,为每一层提供更高级的表征,实现在复杂区域捕获较为显著的物体。

对于背景图像,直接根据全图分割的段数对图像进行排序,返回复杂性得分。对于前景图像,首先判断显著性检测结果,如果图像内显著区域的数量大于 3,则会被认定为过于复杂,将会直接被删除。然后将语义分割得到的结果和显著性检测得到的结果相组合,计算显著区域边缘的分割段数。前景图像的复杂性检测效果如图 2 所示。在这种情况下,若图像内任意一个显著区域边缘段数大于 20,系统认为该图像背景复杂,直接丢弃。

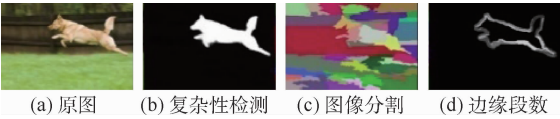


图 2 复杂性过滤过程

Fig. 2 Process of complexity filtering

对于前景图像而言,每个图像中可能包含一个或多个对象,因此对图像全局提取特征是不合适的。如果只对图像的目标区域提取内容特征,便可获取更为准确的特征描述。系统使用 GrabCut 来实现图像目标区域和背景的二值分割。GrabCut 是以 GraphCut 图分割算法为基础改进而得到的交互式图像分割算法,其在分割初始化时加入了大量的用户先验知识和主观需求,从而实现更加准确的目标区域分割效果。图 3 展示了 GrabCut 的基本用法。用户需要使用矩形框手动标记前景区域即可实现图像分割,在分割的过程中,用户可根据主观需求对前景和背景像素进行额外的标注,以改善分割结果。

GrabCut 采取 RGB 颜色空间,通过高斯混合

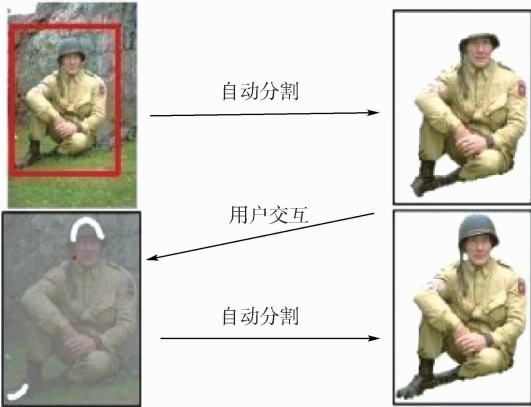


图 3 GrabCut 算法交互过程

Fig. 3 Interaction process of GrabCut algorithm

模型对图像前景及图像背景中的像素进行建模。GrabCut 算法用于分割的 Gibbs 能量函数定义为 $E(\alpha, k, \theta, z) = U(\alpha, k, \theta, z) + V(\alpha, z)$ (1) 式中: α 为透明度系数,当像素为前景时取值 1,为背景时取值 0; $k \in \{1, 2, \dots, k\}$, k 为 GMM 中高斯分量的个数; $\theta = \{\pi_k, u_k, \Sigma_k\}$, 元素分别为每个高斯分量对应的权重系数、均值和协方差; U 和 V 分别表示区域项和能量边界惩罚,计算式如下:

$$U(\alpha, k, \theta, z) = \sum_n D(\alpha_n, k_n, \theta, z_n) \quad (2)$$

$$D(\alpha_n, k_n, \theta, z_n) = -\log \pi(\alpha_n, k_n) + \frac{1}{2} \log \det \Sigma(\alpha_n, k_n) + \frac{1}{2} [z_n - \mu(\alpha_n, k_n)]^T \cdot \Sigma(\alpha_n, k_n)^{-1} [z_n - \mu(\alpha_n, k_n)] \quad (3)$$

$$\theta = \{\pi(\alpha, k), \mu(\alpha, k), \Sigma(\alpha, k), \alpha = 0, 1, k = 1, 2, \dots, k\} \quad (4)$$

$$V(\alpha, z) = \gamma \sum_{(m, n) \in C} [\alpha_m \neq \alpha_n] \exp(-\beta \|z_m - z_n\|^2) \quad (5)$$

式中:参数 $\gamma = 50$; β 由图像的对比度决定;区域项 U 用于表示透明度系数 α 对于像素的相似程度,可由所有像素 $D(\alpha_n, k_n, \theta, z_n)$ 求和得到, $D(\alpha_n, k_n, \theta, z_n)$ 由式(3)计算得到,具体表示为该像素属于前景或背景的概率负对数;边界惩罚项 V 表示为邻域像素 m 和 n 之间由于不连续所带来的惩罚,通过像素 m 与 n 之间的欧氏距离,来对像素之间的相似性进行衡量,如果两邻域像素的相似性越大,说明邻域像素处于同一背景或同一前景的概率越大,所对应的边界项表示的能量值则越小。通过不断进行分割估计和模型参数的学习,将 Gibbs 能量函数 E 降低到最小值实现最佳分割。

GrabCut 是一种和用户可以交互的图像分割算法,考虑到该算法无法满足实时性要求比较高的场合,本文将显著性检测和 GrabCut 结合来实现图像自动分割^[20]。首先,对原始图像进行显著性检测,通过设定适当的阈值获取显著区域二值图像。对于背景简单的图像,显著性检测的结果往往聚集在目标前景区域上,这也是使用复杂性检测过滤图像的目标之一。然后,使用图像的显著区域对 GrabCut 算法进行初始化,并重复迭代运行分割算法,实现目标区域的分割。基于 GrabCut 的自动分割算法训练过程可以总结如下:

- 1) 使用显著性检测得到目标前景区域,并将

目标区域形态学扩张 10 次。

2) 使用目标区域对高斯混合模型进行初始化,目标区域外的其他区域设定为背景区域,区域内设置为待定的未知区域。

3) 使用 K-means 算法将标定的图像前景和图像背景像素聚为 k 类, k 为 GMM 中高斯分量个数, 一般情况取 5, 初始化 GMM 参数 $\theta_k = \{\pi_k, u_k, \Sigma_k\}$ 的值。

4) 以最小化 $D(\alpha_n, k_n, \theta, z_n)$ 的规则判断每个像素所属的 GMM 分量。

5) 根据式(2)计算区域项 $U(\alpha, k, \theta, z)$ 。

6) 根据式(5)计算边界项惩罚 $V(\alpha, z)$, 其中参数 $\gamma = 50, \beta$ 由图像的对比度决定。

7) 构建 s - t 图, 对图像使用最大流、最小割算法进行处理, 得到能量函数 E 最小峰值。

8) 迭代执行步骤 3) 和步骤 7) 直至收敛。

9) 使用 bordermatting 平滑图像分割后的边缘, 这样就实现了结合显著性检测和 GrabCut 的图像目标区域自动提取, 目标区域的提取结果将会同源图像一并返回给用户, 用户可以通过手动指定背景及前景像素的方式来修正图像提取结果。

GrabCut 算法伪代码如下:

- 1: 输入图像 $\{I^1, I^2, \dots, I^H\}$
- 2: AimReg: rect $\leftarrow$ Significance detection method
- 3: Repeat:
- 4: do:
- 5: class $k \leftarrow$ clustering(foreground & background)
- 6: GMM $\leftarrow$ initialize
- 7: Reg: $U(\alpha, k, \theta, z)$, Bp: $V(\alpha, z) \leftarrow$ calculate
- 8: cut_graph $\leftarrow$ maximum_flow(gr, source, sink)
- 9: $E \leftarrow$ mix
- 10: while: 函数收敛或达到迭代次数
- 11: end 循环
- 12: 使用 bordermatting 平滑图像分割后的边缘
- 13: 输出结果

2.2 一致性排序算法设计

在一致性排序算法设计中,检索结果的准确性取决于对特征的精确描述,特征提取的重点在于全面性和准确性。针对不同的应用场景,提出了以下 3 种图像特征提取方法:HSV 颜色直方图特征、HOG 特征及人像姿态特征。

2.2.1 HSV 颜色直方图

本文基于 HSV 色彩空间建立颜色直方图。如图 4 所示,HSV 颜色空间可以通过一个倒立的圆锥体来形象地解释说明,其包含感知三要素:

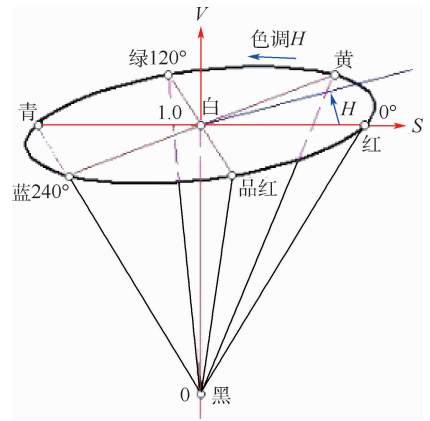


图 4 HSV 颜色空间

Fig. 4 Space of HSV color

色调(H)、饱和度(S)、亮度(V)。亮度对应图中长轴,从黑到白顺序表示像素明暗度;饱和度对应离开长轴的距离,是颜色的深淡度的体现;色调对应围绕长轴的角度,是影响人的视觉判断重要元素。

颜色直方图是对图像的不同颜色区间统计的像素数量分布^[21],其不考虑像素颜色的所处位置,仅与颜色在整幅图像中所占的比例相关。颜色直方图方法可定义为

$$H(k) = \frac{n_k}{N} \quad k = 0, 1, \dots, L - 1 \tag{6}$$

式中: N 为在图像内的所有像素点总数量; n_k 为相同区间内的像素数量。使用 HSV 颜色直方图量化颜色空间可以按照以下步骤^[22]:

1) 根据人眼对颜色近似程度的视觉辨别能力,将色调划分为 8 个不同的区间,饱和度和亮度空间分别划分为 3 个量化区间。

2) 根据像素所属色彩的范围对图像进行量化,此时每个小区间成为直方图的一个簇,这样色彩空间被量化成有限的离散级数。

3) 按照转换式(7)将颜色空间三要素对应的向量参数合并至一维特征向量之中。

$$L = HQ_s Q_v + Q_v S + V \tag{7}$$

式中: Q_s 和 Q_v 分别为分量 S 和 V 的量化级数,本文 $Q_s = 3, Q_v = 3$ 。

式(7)即变化为

$$L = 9H + 3S + V \tag{8}$$

构成了 72 维向量的一维直方图: $L[0, 1, \dots, 71]$ 。之后进行归一化处理,就可以通过向量 L 来衡量图像的颜色特征。

2.2.2 HOG 特征

梯度方向直方图(histogram of oriented gradients, HOG)算子是 Dalal 和 Triggs^[23]提出的一种

基于统计图像局部梯度走向和梯度强度分布的特征描述符。HOG 核心思想是:图像中目标物体边缘位置通常能够直观地反映像素梯度。利用梯度做直方图来进行信息的统计,这种方法能够直观地对区域目标的外在特征进行合理的反映。以 64×64 像素的图像 I 为例,对应的 HOG 特征提取步骤如下^[24-25]:

1) 图像分割。将初始图像等分为 8×8 的 cell,并把 4 个相邻 2×2 的 cell 组成 block。

2) 色彩和伽马归一化。采用 gamma 校正方法对图像进行归一化,该操作能减低光线变化和图像的局部阴影所造成的影响。

3) 计算每个像素的梯度方向和大小。该方法能有效地捕获了图像的轮廓和纹理信息,同时进一步弱化光照、阴影等因素的干扰。图像中像素点 (x,y) 的横向梯度 $H(x,y)$ 及纵向梯度 $V(x,y)$ 计算如下:

$$H(x,y) = I(x+1,y) - I(x-1,y)$$

(9)

$$V(x,y) = I(x,y+1) - I(x,y-1)$$

(10)

该点梯度大小 $G(x,y)$ 和梯度方向 $\theta(x,y)$ 分别为

$$G(x,y) = \sqrt{(H(x,y))^2 + (V(x,y))^2}$$

(11)

$$\theta(x,y) = \arctan\left(\frac{V(x,y)}{H(x,y)}\right)$$

(12)

4) 根据每个像素点的梯度方向,利用双线性内插法将 cell 内像素幅值累加到直方图中,并以 block 为单位把特征向量首尾相连。当把梯度方向均分为 9 块,即可得到一个 64×9 维的 HOG 特征向量。

2.2.3 人像姿态特征

人像姿态识别对于描述人体姿态、区分人体行为至关重要。姿态估计使用人体骨架的形式对一个人的体态进行描述。骨架是一组坐标点,通过对躯干、头部等的关键点识别,将其连接起来用于描述人物的姿态。为了提升人像图像的检索效率,本文使用开源人体姿态识别项目 OpenPose^[26-28] 进行二维姿态检测。OpenPose 是基于深度学习开发的开源库,可以实现对人脸关键点、人手的关键点及人身体的主要关节的实时多人定位。图 5 为人体关键点示意图。每个人像的二维的姿态特征可由 36 维特征向量表示:

$$BP = [(x_A, y_A), (x_B, y_B), \dots, (x_O, y_O)]$$

(13)

2.2.4 基于 Meanshift 的一致性排序

为了提高检索的精度和准确性目标,通过对图片内容一致性重新排序来达到目标。不同于常规的基于内容的图像检索方法,本文通过对候选

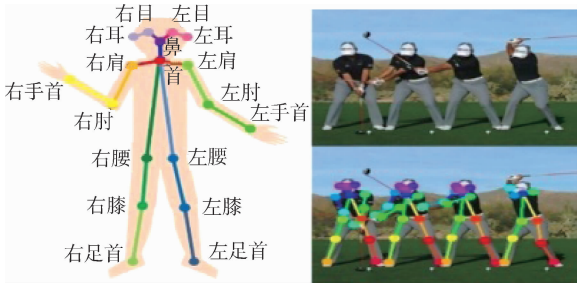


图 5 人像姿态特征
Fig.5 Feature of human pose

图像进行聚类,使用簇中心代替示例图像,用户无需额外提供任何示例图像,即可通过关键字标签实现图像检索。

系统选择均值漂移 (Meanshift) 聚类算法进行聚类操作。Meanshift 算法^[29] 是局部密度最大值的迭代梯度上升法,使用均值漂移向量代表样本中心点局部密度增长的最快方向,将收敛于同个局部密度最大处的样本划分为同一簇类实现数据聚类。Meanshift 算法先在特征空间中随机初始化超球面,在每次迭代中,超球面都朝着均值偏移矢量的方向移动,移动过程中,所有出现在超球面内的数据属于此类的访问频率加 1。均值偏移矢量的计算如下:

$$y_{i+1} = \frac{1}{\Theta} \sum_{\lambda x \in \Theta} (y_i - x)$$

(14)

式中: y_i 为第 i 次迭代的球面中心; x 为属于集合内 Θ 半径为 λ 的球面中的特征点。图 6 表示了 Meanshift 算法超球面初始化到收敛的过程。Meanshift 算法通过引入窗口半径 λ , 避免事先定义聚类数量 k , 故 Meanshift 算法仅需定义参数 λ 即可实现聚类。超球面最终将收敛于局部密度极大值点,并通过判断当前簇中心在阈值范围内是否已经存在其他簇中心确定数据所属簇类,如果存在其他簇,将 2 类数据合并,否则当前簇中心成为新的聚类。当所有数据均被标记访问时算法停止,并按照簇类的访问频率划分数据归属簇类。

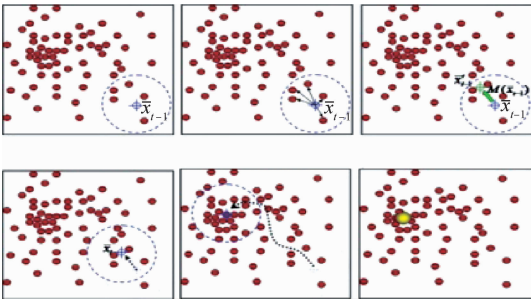


图 6 Meanshift 算法迭代过程
Fig.6 Iterative process of Meanshift algorithm

通过 Meanshift 算法得到最大簇后,计算得到簇中心。按照簇内图像与簇中心之间的距离排序。对于颜色直方图而言,使用卡方距离相较于欧氏距离能够获得更好的检索效果。故针对颜色特征使用卡方距离重排序,卡方距离为

$$D_{\text{Chi}}(x_i, x_j) = \frac{1}{2} \sum_{k=1}^n \frac{(x_{i,k} - x_{j,k})^2}{x_{i,k} + x_{j,k}} \quad (15)$$

对其他特征使用欧氏距离排序,计算如下:

$$D_{\text{Euc}}(x_i, x_j) = 2 \sqrt{\sum_{k=1}^n (x_{i,k} - x_{j,k})^2} \quad (16)$$

使用对应距离公式来表示内容一致性评分,将一致性得分标准化为介于(0,1)之间的值。这样每张图像都被赋予对应特征下的一致性得分。根据不同情况获取对应特征下的分数:对于背景图像使用颜色特征获取一致性得分,并与复杂性得分相结合作为背景图像排序的最终评分依据;对于前景图像同时建立 HSV 颜色直方图特征和 HOG 特征,如果目标对象是人物还将额外使用姿态特征,将图像特征对应的一致性分数线性组合,对数据集内图像进行一致性排序,按照排序结果返回图像,这样就实现了“从无到有”的图像检索。

2.3 图像融合算法设计

图像的无缝融合应该尽可能地保证合成的图像真实自然,减少明显的拼接痕迹。本文提供了 Alpha 融合和泊松图像编辑 2 种无缝融合方案供用户选择。

2.3.1 Alpha 融合算法

Alpha 融合是按照 α 向量(图像透明度)来混合前景像素和背景像素的图像处理技术。首先,把背景像素 B_p 和前景像素 F_p 按照 RGB 三个颜色通道分离,分别对各个颜色进行以下处理:将前景像素乘上 α 的值,同时把背景像素与 α 的反值相乘。然后,在对应的颜色通道上,将处理结果相叠加。最后,把求得的叠加结果除以 α 的最大值获取融合图像的像素值 C_p :

$$C_p = \alpha_p F_p + (1 - \alpha_p) B_p \quad (17)$$

Alpha 融合算法的效果很大程度上取决于前景图像的掩膜 mask 值,当前景图像的掩膜不够精确时,会导致前景图像边缘处的错误结果。此外,Alpha 融合算法是在像素级别对图像进行操作,无法处理图像之间的色彩和纹理变化。当背景与前景图像的纹理差别很大,以及在不同的光照条件下,往往会出现明显的边界,产生不真实融合效果。

2.3.2 泊松图像编辑

泊松图像编辑是在梯度域求解前景和背景重叠区域的像素,利用前景区域的梯度和背景图像

的边界值构造求解带 Dirichlet 边界条件的泊松方程,利用拉普拉斯算子插值重构重叠部分的像素。如图 7 所示, B 和 Ω 分别代表背景图像和背景中的重叠区域,泊松图像编辑的目标是将前景图像的目标区域 U 无缝融合至背景 B 中,即求解融合后重叠区域 Ω 的像素值 f 。通过使边界 $\partial\Omega$ 处梯度变换尽可能小,避免因过渡不自然而出现明显的融合间隙,实现真实自然的融合效果。

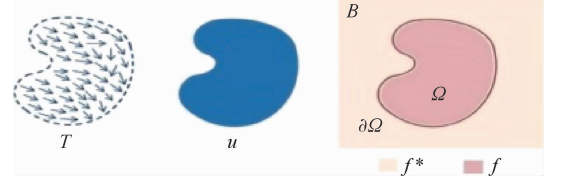


图 7 泊松图像编辑原理

Fig. 7 Schematic diagram of Poisson image editing

为了保证图像融合后重叠区域内变化平滑及无明显的可见边界,重叠区域 Ω 内梯度变换应尽可能得平缓,即梯度应尽可能得小。因此, Ω 内的像素值 f 应满足:

$$\begin{cases} \min_f \iint_{\Omega} \|\nabla f\|^2 \\ f|_{\partial\Omega} = f^*|_{\partial\Omega} \end{cases} \quad (18)$$

式中: f^* 为背景 B 中像素; $\nabla \cdot = \left[\frac{\partial \cdot}{\partial x}, \frac{\partial \cdot}{\partial y} \right]$ 为梯度算子。求解 Ω 内的像素值 f 必然满足关联的欧拉-拉格朗日方程:

$$\begin{cases} \Delta f = 0 \\ f|_{\partial\Omega} = f^*|_{\partial\Omega} \end{cases} \quad (19)$$

式中: Δ 表示拉普拉斯算子。为了保持原图像的纹理特征,需要将梯度矢量场 T 加入到式(19)中,对当前的梯度域加以限制,这样就得到了扩展的最小值问题。公式更新为

$$\begin{cases} \min_f \iint_{\Omega} \|\nabla f - T\|^2 \\ f|_{\partial\Omega} = f^*|_{\partial\Omega} \end{cases} \quad (20)$$

最优解满足 Dirichlet 边界条件下的泊松方程:

$$\begin{cases} \Delta f = \text{div } T \\ f|_{\partial\Omega} = f^*|_{\partial\Omega} \end{cases} \quad (21)$$

$$\text{div } T = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \quad (22)$$

式(22)是 $T = (u, v)$ 的离散形式。在彩色图像的泊松编辑方程求解中,合成图像的颜色在不同通道将被独立构成一个泊松方程求解。 Ω 的边界满足:

$$\partial\Omega = \{p \in B \setminus \Omega : N_p \cap \Omega \neq \emptyset\} \quad (23)$$

式中: N_p 为背景内像素 p 的四连通邻居集; $\langle p, q \rangle$

为 p 和 N_p 内 q 的像素对。泊松方程就转化为了离散的二次最优化问题,如下:

$$\begin{cases} \min \sum_{f|_{\Omega} < p, q > \cap \Omega \neq \emptyset} (f_p - f_q - v_{pq}) \\ f|_{\partial\Omega} = *|_{\partial\Omega} \quad p \in \partial\Omega \end{cases} \quad (24)$$

式中: v_{pq} 为 T 在 $[p, q]$ 上的投影。根据最优性条件,其最优解满足:

$$|N_p| f_p - \sum_{q \in N_p \cap \Omega} f_q = \sum_{q \in N_p \cap \partial\Omega} f_q^* + \sum_{q \in N_p} v_{pq} \quad (25)$$

通过对线性方程求得的解为目标区域内的像素值,这样就实现了图像的无缝融合。

3 基于双向规则的人像布局推荐

同类型的电影海报具有相似的人像分布规律。基于这种考虑,本文建立基于统计的正规则,即从构图良好的图像中提取共通的构图规律。正规则建立的过程可概括为海报分类、使用高斯混合模型对各类别下人像分布建模和高斯混合模型参数估计 3 部分。

3.1 基于背景的海报聚类

先对海报进行聚类,再对不同类别的人像进行单独的位置建模,提取电影海报的布局分布规律。在此,考虑了多个方面的分类准则,如不同电影类别的海报应该具备独特的风格,大众审美会随着时间的推移发生变化,电影海报也会随着改变。电影的类别和时间线都可以作为电影分类的标准,但不论是电影类别的差异还是艺术认知的改变,人像分布和背景图像的内容特征都是紧密相关的。海报制作者在进行人像布局设计时,前景人像与背景图像之间的协调搭配都是应该考虑的第一要素。故选择依据图像背景内容特征对电影海报进行分类。首先,切除海报内人像和其他前景元素,使用内容感知算法填充背景上的空白区域。然后,建立背景图像的内容特征。考虑到有些海报内可能没有明显的显著元素,故考虑依据图像颜色特征进行分类。建立背景的 HSV 颜色直方图特征,并使用 K-means 聚类算法^[30]对填充的背景分类。该算法的一个特征是每次迭代都需要计算所有样本与质心之间的相似度,在数据规模较大时,算法的时间开销比较大。

3.2 高斯混合模型构建

海报分类完成后,分别对每类背景的人像布局关系进行分析,在这里同时考虑人像的位置坐标和尺寸。建立一个生成概率模型来训练每个场景类别的位置、大小分布,使用高斯混合模型来拟合人像布局分布。高斯混合模型作为广泛应用的

聚类模型之一,使用多个高斯分布作为参数模型,并将其线性组合来刻画数据的一般分布。对于任何场景类别 I ,本文将人像布局信息表示为 $x(I) = (x, y, r)^T$, 其中 (x, y) 表示人物在图像中的平均位置, r 表示人像的面部尺寸。给定场景 I_k 的人像布局 $x(I_k)$ 的概率分布可以表示为

$$p(x(I_k)) = \sum_{i=1}^{N_k} w_i^k N(x | \mu_{i,k}, \Sigma_{i,k}) \quad (26)$$

$$N(x | \mu_{i,k}^x, \Sigma_{i,k}^x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_{i,k}^x|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \mu_{i,k}^x)^T \Sigma_{i,k}^{x-1} (x - \mu_{i,k}^x)} \quad (27)$$

$$\sum_{i=1}^{N_k} \omega_i = 1 \quad 0 \leq \omega_i \leq 1 \quad (28)$$

式中: $N(x | \mu, \Sigma)$ 为高斯概率密度函数; w_k, μ_k, Σ_k 分别为混合模型中的第 k 维高斯向量的权重、均值、协方差矩阵。

3.3 高斯混合模型参数估计

本文使用贝叶斯信息准则^[31] (Bayesian information criterion) 来估计高斯混合分量的数量。使用最大期望 (expectation-maximization, EM) 算法^[32] 估计所有背景类别的高斯混合模型参数 (w_k, μ_k, Σ_k) 。EM 算法是通过迭代进行极大似然估计 (maximum likelihood estimation) 的优化算法,通常用于对包含隐变量或缺失数据的概率模型进行参数估计。EM 算法包括 2 个步骤: E-step 和 M-step。在 E-step 计算中,通过评估对隐藏变量的现有估计值,计算 E-step 中最大似然估计值:

$$p(i | x_t) = \frac{N(x_t | \mu_i^k, \Sigma_i^k) w_i^k}{\sum_{i=1}^{N_k} N(x_t | \mu_i^k, \Sigma_i^k) w_i^k} \quad (29)$$

在 M-step 计算中,通过把 E-step 上求得的最大似然值最大化操作后,来计算参数的值。每个类的潜在变量平均值 μ_k 、协方差 Σ_k 和先验值 w_k 更新如下:

$$\hat{\mu}_i^k = \frac{\sum_{t=1}^{T_k} p(i | x_t) x_t}{\sum_{t=1}^{T_k} p(i | x_t)} \quad (30)$$

$$\hat{\Sigma}_i^k = \frac{\sum_{t=1}^{T_k} p(i | x_t) [x_t - \hat{\mu}_i^k] [x_t - \hat{\mu}_i^k]^T}{\sum_{t=1}^{T_k} p(i | x_t)} \quad (31)$$

$$\hat{w}_i^k = \frac{\sum_{t=1}^{T_k} p(i | x_t)}{T_k} \quad (32)$$

式中: T_k 为给定场景类别 I_k 的海报总数量。M-

step 计算更新的参数结果将被重新应用于 E-step 的计算中,2 个步骤迭代的交替运行,模型收敛时的结果即可表示人像概率分布。

3.4 基于美学的负规则

不同的背景图像具备独特的特点,这些特点决定了不适合人像的位置区域。对于不同前景和背景的组合,人像布局推荐应该随之改变。本文受到摄影中美学常识的启发,总结建立了 3 条负规则^[33]。

1) 区域规则。指人像图像应尽可能少地覆盖其他前景元素和背景中的显著区域。区域规则确保了在构图设计时的前景元素间主次分明、简洁明了。如果前景元素间出现相互覆盖的情况,图像可能会出现视觉上的逻辑错误;而背景中的显著区域往往是相对重要的部分,表现了图像的视觉特点,人像与背景显著区域保留一定的空间。首先,使用本文设计的显著性检测方法检测图中显著区域。然后,通过式(33)量化显著区域负规则以评估显著区域与人像布局间的合理性。

Score⁻(W/M) = 1 - λ_M $\frac{\|Rect_w \& M\|}{\|M\|}$ (33)

式中:M 为显著区域的二元掩码,可由 3.1 节的方法生成,或者由用户手动标注;Rect_w为在位置 W 处放置时对象所占据的矩形区域;符号 & 为与操作,该操作计算对象所覆盖的突出区域。

2) 直线规则。直线规则是指背景内的直线不应穿透人像的头部。画面中线条是构图的骨架,是整个画面造型的经络。画面中线条结构的走势能够很好地衬托照片内容的表现,起到美化照片的作用。本文采用霍夫变换(Hough transform)^[34]检测背景中的长直线。霍夫变化被认为是目前最通用也是效果最好的直线检测算法。

图 8 展示了霍夫直线检测流程。首先,将图像中的所有像素点(坐标值)投影到参数坐标系下;然后,在参数坐标系中搜索极值点,确定代表直线的峰值点个数和坐标。当候选的线段被识别出来时,将归属于同一条直线的线段相连接,检测直线的起始点和终点。这样就实现了霍夫变换下的直线检测:

Score⁻(W|L) = 1 - λ_L × 1(Head_w & L) (34)

3) 灭点规则。灭点规则是指人像主体不应该覆盖背景的灭点。在线性延展中,从起点出发,如果多条平行线向远处延展时,最终总是近似相交于一点,该点称之为灭点。灭点反映了图像深度的渐变和透视表现力,如果背景出现灭点,应尽量减少人像源点和背景灭点间的相互覆盖,避免

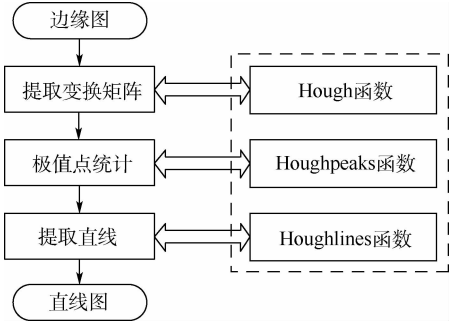


图 8 霍夫直线检测流程

Fig. 8 Process of Hough line detection

影响以视觉感官为线性走势的空间次序。

二维空间中的灭点概念如图 9 所示,场景中的每一个灭点都有其对应的灭线,这样灭点检测就能转换为同组的直线检测。本文在进行灭点检测时,采用了 J-Linkage 算法。J-Linkage 算法最初由 Toldo 和 Fusiello^[35]提出,属于多模型点聚类算法,是现阶段灭点检测中最优的算法。算法核心在于构造优先集矩阵及对矩阵的优化计算。优先集矩阵一般为 N × M 的布尔矩阵,假定有 N 条直线集合和 M 个灭点集合。优先级矩阵描述了每条直线和每个假设灭点之间的连续性关系。优先级矩阵优化迭代的过程就是对直线分类的过程。首先,判断不同直线间的优先级是否相交,将相交的直线直接划分为同类直线。然后,使用 Jaccard 距离合并不同类别下的直线。Jaccard 距离可用于判断 2 个集合间的差异性,表达式如下:

D_{Jac}(A,B) = $\frac{|A \cup B| - |A \cap B|}{|A \cup B|}$ (35)

式中:A、B 为 2 个不同的直线集合。通过计算优先集矩阵内所有直线集合间的 Jaccard 距离,将 Jaccard 距离最小的集合组合为同一集合。重复迭代上述操作,直到优先集矩阵内所有直线集合间的 Jaccard 距离均为 1。这种情况下,所有直线集合间的差异化达到最大,也就实现了直线间的分类。

Score⁻(W|Z) = 1 - λ_Z × 1(Rect_w(Z) = 1) (36)

通过式(36)量化灭点规则以评估显著区域与人像布局间合理性。其中,Z 表示灭点的二维坐标。

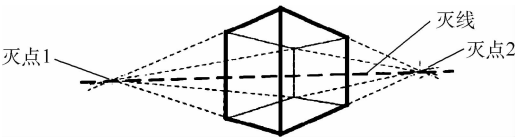


图 9 灭点概念

Fig. 9 Concept of vanishing point

3.5 双向规则的组合

对于用户自己确定的背景图像,首先,提取图像像的 HSV 颜色直方图特征 L ,通过比较特征 L 与背景簇中心 I_k ,选择距离最小的簇 I_k 作为该图像类别,根据该类别下的人像分布概率,根据式(37)计算出正规则得分。然后,检测图像内显著区域、直线和灭点,根据式(38)将 3 条负向规则组合构成负规则得分。

$$\text{Score}^+(H) = p(H | I_k) \tag{37}$$

$$\text{Score}^-(H) = \text{Score}^-(H | S) \cdot \text{Score}^-(H | V) \cdot \text{Score}^-(H | L) \tag{38}$$

最终将正负规则结合起来,即可获得该背景下的人像布局得分:

$$\text{Score}(H) = \text{Score}^+(H) \times \text{Score}^-(H) \tag{39}$$

只有当一个位置同时满足正规则和负规则,该位置才会获得一个相对较高的双向规则总分,这个位置才会被推荐给用户。

4 实验与分析

4.1 基于关键字的检索与融合实验

4.1.1 图像检索效果测试

从互联网社交媒体中准确地发现图像是一个极富挑战性的研究热点。本文为了对系统进行有效的评估,采取人工分别对互联网直接返回的结果、复杂性方案过滤后的结果及内容一致性排序后的结果的假阳率(false positive rate, FPR)进行评估。假阳率计算式为

$$\text{FPR} = \frac{|\{ \text{FP} \}|}{|\{ \text{FP} + \text{TN} \}|} \tag{40}$$

假阳率表示为负样本被识别为真样本事件发生的概率。为了验证基于互联网的图像检索方法的有效性,使用几个例子测试了图像检索系统。图 10 展示了本文方法检索的结果。本文的图像均来源于互联网搜索引擎,对于每个关键字标签,使用 Google Image 自动下载 500 张图像,经过复



图 10 图像检索结果

Fig. 10 Results of image retrieval

杂性过滤、一致性排序后返回前 100 张图像作为图像素材用于后续图像合成。如表 1 所示,在使用搜索引擎返回的前 100 张图像中,假阳率平均约为 81%,即只有约 19% 的图像是符合目标预期的。经过图像复杂性过滤、和一致性排序后,假阳率分别降低至 71%、26%。

表 1 图像检索结果

Table 1 Results of image retrieval

检索目标	假阳率/%					
	Park	Jump dog	Polar bear	Soccer player	Goal keeper	Manthrow
互联网检索	94	80	78	86	80	71
复杂性过滤	78	76	70	61	76	65
一致性排序	34	27	28	30	18	21

对于前景图像,观察到复杂性过滤主要排除了一些拼接的图像和少量背景复杂的图像,对于提升准确率而言表现的并不出色。然而,它对于一致性排序的提升却非常重要,因为它保证图像在自动分析时可以获得良好的分割效果,这对于内容一致性排序的提升是非常显著的。由于本文加入了姿态估计,在检索具有特定动作的人物,如“Goalkeeper”时检索效果也有所提升。另外,在使用互联网引擎进行图像检索时,适当的对关键字标签添加适当的动词或并列词,如“Manthrow frisbee”和“Jump dog”来获取目标图片是非常有帮助的。由于所有数据是依托互联网搜索引擎获取的,系统依然受制于初步检索的结果。例如,在检索某类人物时,将关键字标签定义为“Prince”,检索结果多是明星或现代皇室成员,这显然与用户预期不同。因此,推荐用户在检索过程中同时添加多个并列的关键字标签或根据返回的结果交互性地修正关键字标签来提高检索的准确率。

4.1.2 图像无缝融合效果测试

本节主要对图像融合算法的效果进行测试。图像合成包括图像目标区域的分割和无缝融合 2 步,本文提供了自动的图像提取,系统根据抓取结果主动生成前景掩码。如果用户对自动提取的结果不满意,可手动实现图像抓取。

图 11 为 Cheetah 和 Polar bear 的源图像、Alpha 融合结果和泊松图像编辑的融合效果。可以发现,在面对不同的前景、背景组合时,2 个算法都表现出了各自的特点。Alpha 融合能保证源图像本身的特点不被改变,在图像的纹理或光照条件相差较大,会产生不真实的融合效果。泊松图像编辑能保持前景和背景之间保持一致的纹理及光照条件,这会导致图像发生失真的现象。在



图 11 图像无缝融合结果

Fig. 11 Results of seamless integration of image

给出的2例融合结果中,采用的不同融合算法各有各自的优势,用户可以根据不同需求选择适合的场景进行融合,从而达到更好的融合效果。

4.2 基于双向规则的人像布局推荐实验

4.2.1 人像布局推荐效果测试

在图像合成中,进行人像布局推荐是新研究的问题,采用了基于双重规则的人像布局推荐系统对背景图像生成人像布局推荐。首先,使用TMDB 官方 API 收集海报 10 000 张,共收集以人像为主题的海报 3 574 张。然后,提取图像背景的 HSV 颜色直方图特征,并使用 K-means 将海报分为 8 类,利用高斯混合模型进行建模,对应类别下的人像概率分布如图 12 所示。

根据 3 条反向规则,检测背景的显著区域、直线和灭点。图 13 展示了显著区域、直线和灭点的检测结果。需要说明的是,系统提供了手动标注的方式来修正检测结果。



图 12 人像分布

Fig. 12 Distribution of human position

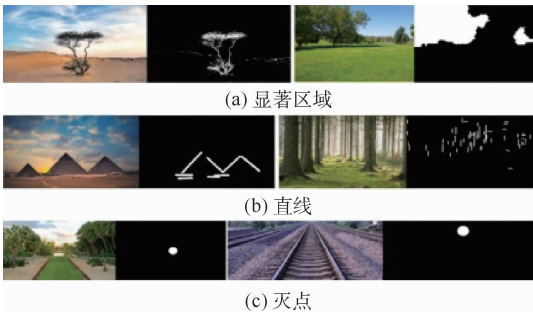


图 13 负规则检测结果

Fig. 13 Results of negative rule detection

将双向规则的最终结果得分图相加即可得到最终的推荐结果。目前,没有明确的指标衡量人像推荐结果,只能通过人工来评判推荐结果的合理性。系统将人像推荐的结果直接应用于图像合

成中,辅助用户进行图像布局设计。

图 14 展示了正规则、负规则及双向规则组合下的推荐结果。默认情况下,系统会根据给定的前景人像关键字标签数量,自动实现对应的人像布局推荐。用户可在手动地添加非人像元素或对人像元素位置进行自定义的修改。每当一张前景图像加入到背景内,背景图像下的该区域将被定义为显著区域。这样就避免了人像之间与物体之间的相互重叠,预防最终合成结果由于图像间的覆盖而出现的逻辑错误。



图 14 人像布局推荐结果

Fig. 14 Results of human position recommendation

4.2.2 系统整体测试

将图像检索、人像布局推荐和图像融合组合在一起,对整个系统的使用效果进行了测试。对系统的多个应用场景示例进行测试。最终生成结果如图 15 所示。①用户输入 *Cinderella* 的文本故事,根据关键字筛选结果,使用“Royalprince”和“Bride”作为关键字人物标签,将“Garden”作为背景标签。检索到对应图像后,将所需的物体抠取出来,并通过人像布局推荐结果将目标图像粘贴到了新的背景中。②用户直接输入“Manthrow”和“Jump dog”前景关键字标签,以及“Park”的背景关键字标签。其中,“Manthrow”为前景人物标签,“Jump dog”为普通的前景标签。③用户将“Ski man”设置为前景人物标签,“Polar bear”作为普通的前景标签,“Snow field”作为背景标签。④用户输入 *The Wonderful Wizard of Oz* 的文本故事信息,系统将“Little girl dorothy”、“Tin woodman robot”和“Scarecrow”作为前景关键字,将“Forest”作为背景关键字。其中,将“Little girl dorothy”设定为人物标签,其他均为普通标签。用户需要手动为普通的前景元素定义位置和大小信息,随着普通前景元素的加入,人像布局推荐结果也会随之改变。系统会自适应地为所有的前景人物推荐布局信息,并根据人脸的尺寸自动缩放图像,实现前景人物元素的自动拼接。从上述 4 个例子中可以发现,依靠本文系统能够根据文本信息“从无到有”地实现目标图像合成,并且所合成的图像十分真实自然。



图 15 系统整体结果

Fig. 15 Final result of system

5 结 论

本文提出了一种基于关键字的海报自动合成系统,并设计了图像检索、图像布局及图像融合的具体方案。在图像检索阶段,本文设计了文本和内容的双重过滤方案,该方案采用复杂性过滤和一致性排序的方法实现从海量的图像库中准确快捷地检索图像的目的,降低了检索过程中的假阳性率,为用户提供了精准快捷的图像检索手段。在图片布局设计阶段,本文设计了一种基于双向规则的人像布局推荐方案,该方案采用正负规则组合的方法提升了海报设计中人像布局中前景人像与背景图像之间的协调度。在图片融合阶段,采用显著性检测和 GrabCut 相结合的算法解决了 GrabCut 算法不能实时处理的问题,实现了图像实时高效的自动分割,并且提供了 Alpha 融合结果和泊松图像编辑的融合 2 种融合方法供不同场景选择。

参考文献 (References)

[1] JOHNSON M, BROSTOW G J, SHOTTON J, et al. Semantic photo synthesis[J]. Computer Graphics Forum, 2006, 25 (3) : 407-413.

[2] CHEN T, CHENG M M, TAN P, et al. Sketch2Photo[J]. ACM Transactions on Graphics, 2009, 28 (5) : 1-10.

[3] NIBLACK C W, BARBER R, EQUITZ W, et al. QBIC project: Querying images by content, using color, texture, and shape[C] // Proceedings SPIE 1908, Storage and Retrieval for Image and Video Databases, 1993, 1908: 173-187.

[4] PENTLAND A, PICARD R W, SCLAROFF S. Photobook: Content-based manipulation of image databases[J]. International Journal of Computer Vision, 1996, 18 (3) : 233-254.

[5] AGARWALA A, DONTCHEVA M, AGRAWALA M, et al. Interactive digital photomontage[C] // ACM SIGGRAPH 2004. New York: ACM, 2004: 294-302.

[6] BEZDEK J C. Modified objective function algorithms[M]. Berlin: Springer, 1981: 155-201.

[7] BOYKOV Y Y, JOLLY M P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images[C] // Proceedings 8th IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2001: 105-112.

[8] ROTHER C, KOLMOGOROV V, BLAKE A. GrabCut; Interac-

tive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics, 2004, 23 (3) : 309-314.

[9] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient graph-based image segmentation[J]. International Journal of Computer Vision, 2004, 59 (2) : 167-181.

[10] WANG J, COHEN M F. Simultaneous matting and compositing[C] // 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007: 1-8.

[11] FATTAL R, LISCHINSKI D, WERMAN M. Gradient domain high dynamic range compression[C] // ACM SIGGRAPH 2002. New York: ACM, 2002: 249-256.

[12] PÉREZ P, GANGNET M, BLAKE A. Poisson image editing[C] // ACM SIGGRAPH 2003. New York: ACM, 2003: 313-318.

[13] INABA S, KANEZAKI A, HARADA T. Automatic image synthesis from keywords using scene context[C] // Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014: 1149-1152.

[14] BHATTACHARYA S, SUKTHANKAR R, SHAH M. A framework for photo-quality assessment and enhancement based on visual aesthetics[C] // Proceedings of the 18th ACM International Conference on Multimedia, New York: ACM, 2010: 271-280.

[15] ZHANG Y H, SUN X S, YAO H X, et al. Aesthetic composition representation for portrait photographing recommendation[C] // 2012 19th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2012: 2753-2756.

[16] WANG Y T, SONG M L, TAO D C, et al. Where2Stand[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 7 (1) : 1-22.

[17] RAWAT Y S, SONG M L, KANKANHALLI M S. A spring-electric graph model for socialized group photography[J]. IEEE Transactions on Multimedia, 2018, 20 (3) : 754-766.

[18] YANG Z, LEI J J, FAN K F, et al. Keyword extraction by entropy difference between the intrinsic and extrinsic mode[J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392 (19) : 4523-4531.

[19] HOU Q B, CHENG M M, HU X W, et al. Deeply supervised salient object detection with short connections[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 5300-5309.

[20] CHENG M M, MITRA N J, HUANG X, et al. Global contrast based salient region detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37 (3) : 569-582.

[21] RAO A, SRIHARI R K, ZHANG Z. Spatial color histograms for content-based image retrieval[C] // Proceedings 11th International Conference on Tools with Artificial Intelligence. Piscataway: IEEE Press, 1999: 183-186.

[22] 童振兴. 基于内容的图像检索技术综述与展望[J]. 计算机光盘软件与应用, 2010, 13 (6) : 88.

TONG Z X. Content-based image retrieval technologies summary and prospects[J]. Computer CD Software and Applications, 2010, 13 (6) : 88 (in Chinese).

[23] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005: 886-893.

[24] XIE Y, LIU L F, LI C H, et al. Unifying visual saliency with HOG feature learning for traffic sign detection[C] // 2009 IEEE

Intelligent Vehicles Symposium. Piscataway: IEEE Press, 2009: 24-29.

[25] TERASAWA K, TANAKA Y. Slit style HOG feature for document image word spotting [C] // 2009 10th International Conference on Document Analysis and Recognition. Piscataway: IEEE Press, 2009: 116-120.

[26] WEI S H, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4724-4732.

[27] SIMON T, JOO H, MATTHEWS I, et al. Hand keypoint detection in single images using multiview bootstrapping [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 4645-4653.

[28] CAO Z, SIMON T, WEI S H, et al. Realtime multi-person 2D pose estimation using part affinity fields [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1302-1310.

[29] CHENG Y Z. Mean shift, mode seeking, and clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(8): 790-799.

[30] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient K-means clustering algorithm: Analysis and implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.

[31] SCHWARZ G. Estimating the dimension of a model [J]. The Annals of Statistics, 1978, 6(2): 461-464.

[32] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.

[33] ZUCKER M. Monte Zucker's portrait photography handbook [M]. [S. l.]: Amherst Media, 2007.

[34] BALLARD D H. Generalizing the Hough transform to detect arbitrary shapes [J]. Pattern Recognition, 1981, 13(2): 111-122.

[35] TOLDO R, FUSIELLO A. Robust multiple structures estimation with J-linkage [C] // European Conference on Computer Vision. Berlin: Springer, 2008: 537-547.

Automatic poster synthesis system based on keywords

GUAN Shuaipeng, YU Haiyang, YANG Zhen*, ZHOU Ming, LAI Yingxu

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The popularization of intelligence puts forward new requirements for image editing. As a way of transmitting information in the form of images, posters play an important role in daily life and work management. But the production of posters requires multi-element image synthesis. However, there is a lack of an interactive and one click image synthesis system. Combined with the current popular image processing technology, a poster automatic synthesis system is designed and implemented. We propose keyword-based image retrieval scheme, constructs a dual filtering scheme based on text and content, and provides users with accurate and fast image retrieval methods. By counting the composition rules of a large number of carefully designed poster pictures and introducing the composition rules of aesthetic common sense, we propose a portrait layout recommendation scheme based on two-way rules, which assists users in portrait layout design under the combined effect of two-way rules. The experimental results prove that the scheme designed in this paper can run stably and efficiently, users can realize image synthesis through simple interactive operations, and the final image synthesis effect is real and effective.

Keywords: image retrieval; automatic segmentation; seamless integration; layout recommendation; automatic synthesis

Received: 2020-09-27; **Accepted:** 2020-12-04; **Published online:** 2021-03-18 10:17
URL: kns.cnki.net/kcms/detail/11.2625.V.20210317.1118.001.html
Foundation items: National Natural Science Foundation of China (61671030); Beijing Great Wall Scholar (CIT&TCD20190308); China Postdoctoral Science Foundation (2019M660377)
* **Corresponding author.** E-mail: yangzhen@bjut.edu.cn

《北京航空航天大学学报》征稿简则

《北京航空航天大学学报》是北京航空航天大学主办的以航空航天科学技术为特色的综合性自然科学学术期刊(月刊)。本刊以反映航空航天领域研究成果与动态、促进学术交流、培养科技人才和推动科技成果向社会生产力转化为办刊宗旨。本刊为中国自然科学技术核心期刊,并被 Ei Compendex 等国内外权威文献检索系统收录。本刊向国内外公开发行人,为进一步提高办刊质量和所刊出文章的学术水平,特制定本简则。

1 论文作者及内容

1.1 本刊面向海内外所有学者。

1.2 主要刊载与航空航天科学技术有关的材料科学及工程、飞行器设计与制造、宇航科学与工程、信息与电子技术、控制技术和自动化工程、流体力学和动力工程、计算机科学及应用技术、可靠性工程与失效分析等领域的研究文章。航空航天科学技术民用方面以及具有航空航天工程背景的应用数学、应用物理、应用力学和工程管理等方面的文章也在本刊优先考虑之列。

2 来稿要求

2.1 论文应具有创新性、科学性、学术性和可读性。

2.2 论文为原创作品,尚未公开发表过,并且不涉及泄密问题。若发生侵权或泄密问题,一切责任由作者承担。

2.3 主题明确,数据可靠,图表清晰,逻辑严谨,文字精练,标点符号正确。

2.4 文稿撰写顺序:中文题名(一般不超过 20 个汉字),作者中文姓名、单位、所在城市、邮政编码,中文摘要(包括目的、方法、结果及结论),中文关键词(5~8 个),中图分类号,英文题名,作者英文姓名、单位、所在城市、邮政编码、国别,英文摘要,英文关键词,引言,正文,参考文献。首页下角注明基金项目名称及编号,作者信息。

2.5 作者请登录本刊网页进行在线投稿。

3 稿件的审核、录用与版权

3.1 来稿须经专家两审和主编、编委讨论后决定刊用与否。

3.2 若来稿经审查后认定不宜在本刊发表,将及时告知作者。如果在投稿满 3 个月后仍未收到本刊任何通知,作者有权改投它刊。在此之前,请勿一稿多投,否则一切后果自负。

3.3 来稿一经刊登,即赠送单行本。

3.4 来稿一经作者签字并在本刊刊出,即表明所有作者都已经认可其版权转至本刊编辑部。本刊在与国内外文献数据库或检索系统进行交流及合作时,不再征询作者意见。

邮寄地址:100191 北京市海淀区学院路 37 号 北京航空航天大学学报编辑部

办公地点:北京航空航天大学办公楼 405,407,409 房间

电 话:(010)82315594,82338922,82314839,82315426

E-mail: jbuaa@buaa.edu.cn

http://bhxb.buaa.edu.cn

http://www.buaa.edu.cn

《北京航空航天大学学报》 第五届编辑委员会

主 任 (主 编): 赵沁平

(以下按姓氏笔画为序)

副主任 (副主编): 丁希仑 王少萍 孙志梅 李秋实 李焕喜 杨嘉陵
苗俊刚 相 艳 徐立军 钱德沛 曹晋滨

编 委: 马殿富 王 琪 王 聪 邓小燕 王青云 王荣明 刘 宇
刘 红 江 洁 刘 强 闫 鹏 朱天乐 刘铁钢 齐铂金
陈万春 邹正平 苏东林 杨世春 沈成平 邱志平 宋知人
杨树斌 张晓林 杨晓奕 杨继萍 李惠峰 吴新开 张瑞丰
杨照华 宋凝芳 周 锐 林宇震 林贵平 战 强 姚仰平
胡庆雷 赵秋红 段海滨 赵巍胜 席 平 郭 宏 徐 洁
徐世杰 郭洪波 康 锐 翟 锦 熊华钢

北京航空航天大学学报

Beijing Hangkong Hangtian Daxue Xuebao

(原《北京航空学院学报》)

(月刊 1956 年创刊)

第 48 卷 第 2 期 2022 年 2 月

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUA)

(Monthly, Started in 1956)

Vol.48 No.2 February 2022

主管单位 中华人民共和国工业和信息化部
主办单位 北京航空航天大学
主 编 赵沁平
编辑出版 《北京航空航天大学学报》编辑部
邮 编 100083
地 址 北京市海淀区学院路 37 号
印 刷 北京科信印刷有限公司
发 行 《北京航空航天大学学报》编辑部
发行范围 国内外发行
联系电话 (010) 82315594 82338922
82314839
电子信箱 jbuaa@buaa.edu.cn

Administrated by Ministry of Industry and Information
Technology of the People's Republic of China
Sponsored by Beijing University of Aeronautics
and Astronautics (BUAA)
(Beijing 100083, P. R. China)
Chief Editor ZHAO Qinping
Edited and Published by Editorial Board of JBUA
Printed by Beijing Kexin Printing Co., Ltd.
Distributed by Editorial Board of JBUA
Telephone (010) 82315594 82338922
82314839
E-mail jbuaa@buaa.edu.cn
http://bhxb.buaa.edu.cn

中国标准连续出版物号: ISSN 1001-5965
CN 11-2625/V

国内定价: 50.00 元 / 期

ISSN 1001-5965

