基于组合分类器的恶意域名检测技术

盛剑涛，陈茂飞，刘东鑫，汪来富，史国水，金华敏
（中国电信股份有限公司研究院，广东 广州 510630）

摘　要：域名系统是互联网中不可或缺的关键基础服务，难以避免遭到不法分子的滥用。在研究僵尸网络和DGA恶意域名应用的基础上，比较了当前主流的恶意域名检测技术，提出了基于组合分类器的恶意域名检测技术框架。该技术框架以支持向量机为主分类器，融合了朴素贝叶斯分类器模型和其他统计特征。实验数据表明，该技术框架在离线训练时长、对未知DGA恶意域名家族的检测能力方面表现优秀，可以较好地满足运营商大网环境下对恶意域名的检测分析要求。

关键词：恶意域名；僵尸网络；机器学习；深度学习；组合分类器
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020150

Detection of malicious domain name based on a classifier combination

SHENG Jiantao, CHEN Maofei, LIU Dongxin, WANG Laifu, SHI Guoshui, JIN Huamin
Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: As a fundamental service on the internet, domain name system (DNS) can inevitably be abused by malicious activities. Based on the studies of Botnets and other malwares which made use of the domain generation algorithm (DGA), and researches on current major techniques of malicious domain detection, a malicious domain detection framework based on a classifier combination was proposed. The framework applied the support vector machine (SVM) as its main classifier and combined the naive Bayes classifier (NBC) supportively with some statistical characteristics. Experiment result demonstrates that the framework outperformes current techniques in the offline-training time and the capability of detecting unknow malicious domain families, which satisfies the requirement of internet service provider (ISP) to detect and analyze malicious domainson the internet.

Key words: malicious domain name, Botnet, machine learning, deep learning, classifier combination

1　引言

域名系统（domain name system，DNS）提供域名和IP地址的双向映射，是互联网中不可或缺的一项关键基础服务。众多互联网应用（如Web应用、电子邮件和文件传输等）都广泛地依赖DNS。与此同时，DNS也难以避免地遭到僵尸网络（Botnet）、计算机病毒等恶意程序的利用。例

收稿日期：2020−03−15；修回日期：2020−04−22

如 2017 年肆虐全球的 WannaCry 勒索病毒，在代码中以一个域名的查询结果作为恶意活动的开关：如果能获取该域名的解析 IP 地址响应，则停止恶意传播；否则继续[1]。

近年来，以 Wannacry 为代表的恶意程序“趋利化”明显，以“海莲花”为代表的 APT 攻击朝着长期渗透的方向不断发展[2]。这些攻击行为的关键在于，恶意程序和控制中心之间要建立稳健、隐蔽的通信信道。随着技术的进步，攻击者逐步放弃了早期使用硬编码将通信 IP 地址写入恶意程序的方式，转而利用 DGA（domain generate algorithm，域名生成算法）生成临时域名，实现控制中心和恶意程序的远程通信。DGA 以时间戳、随机数和其他自定义信息等作为输入，辅以攻击者自定义的伪随机化算法，以固定的时间周期生成可预测、碰撞概率极小（如果域名已经被注册，则属于碰撞）的随机字符域名，其原理如图 1 所示。

这种方法给网络安全的检测带来极大挑战，DGA 域名的动态变化、随机性可以轻易绕过防火墙等传统安全防护组件，因为后者通常以固定规则的方式实现对已知威胁的检测和防护。

当前，由于 DGA 域名被广泛应用于诸如僵尸网络、勒索病毒等恶意程序中，分析检测 DGA 域名也成为目前学术界和工业界的研究热点。本文着重针对运营商大网域名解析日志的安全分析场景，提出了基于组合分类器的恶意域名检测技术。该方法深入挖掘 DGA 域名和普通域名的差异，以加快离线训练速度、加强对未知 DGA 域名的检测能力为核心目标，实现对 DGA 域名的高精度、低漏报检测。

2 相关工作

由于 DGA 域名的灵活性及低成本，大量的高级恶意程序为了提高隐蔽性，以实现长期、远程控制的目的，普遍利用 DGA 域名来实现 CC（challenge collapsar）和“肉鸡”服务器的通信连接。对于 DGA 域名的检测识别，本质上是对大量的域名进行模式分类的过程，可看作一个二分类（是/否为 DGA 域名）的分类问题。DGA 域名在形式上符合 IETF RFC 关于域名命名的规范，但是不符合日常的域名构造习惯，因为 DGA 域名的字符、长度随机化等特征，其比 IP 地址更难以记忆。

针对 DGA 域名的检测算法，学术界和工业界已进行了大量研究。目前业界的主流方法是，将域名字符作为短文本，并根据数据源或应用场景的特点，补充相关的特征信息（如 DNS 报文的域名 TTL 值、域名 AS 号等[3]），然后采用有监督学习算法构造分类器，最后实现在线识别。这些方法各有长处，其效果与应用场景密切相关。

2.1 基于朴素贝叶斯的检测技术

朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础以及稳定的分类效率。同时，它所需估计的参数很少，对缺失数据不太敏感，算法过程也比较简单。理论上，朴素贝叶斯与其他分类方法相比具有最小的误差率，但实际上并非总是如此。这是因为朴素贝叶斯模型有一个前提假设，即属性之间相互独立，这个假设在实际应

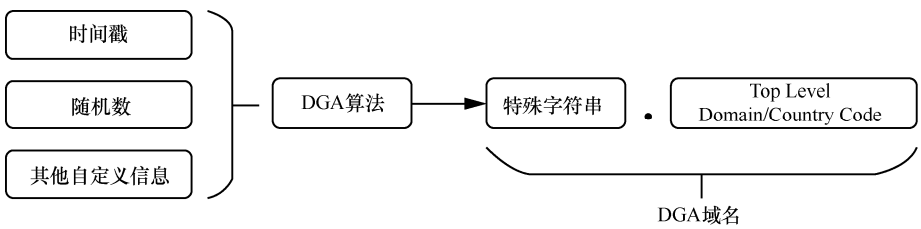


图 1 DGA 域名的生成原理示意图

用中往往是不成立的。此外，朴素贝叶斯模型对输入数据的表达形式很敏感，尤其在融合不同类型特征的时候，在特征抽取方面需要引入额外的处理技巧。这些都给朴素贝叶斯在恶意域名的检测中带来一定的影响。

参考文献[4]提出了以 1-gram、2-gram、3-gram 为域名特征，基于朴素贝叶斯模型构建分类器，以实现对 DGA 域名的有效检测。它的思路以域名的 *n*-gram 字符特征为核心，暂未考虑引入如域名长度、元辅音比例等独立特征。本文认为，仅从域名的 *n*-gram 字符特征出发，模型的检测精度瓶颈明显，并且模型的分类精度容易呈现剧烈震荡的结果。

2.2　基于支持向量机的检测技术

支持向量机（SVM）模型基于 AT&T Bell 实验室提出的向量统计学习理论，是结构风险最小化原理的近似实现，具有高泛化性能的通用学习机器。它是一种有坚实理论基础的小样本学习方法，避开了从归纳到演绎的传统过程，实现了从训练样本到预报样本的高效的“转导推理”。总体来看，支持向量机的训练速度快，模型稳健性好，也能很好地融合不同类型的特征，但是在用于恶意域名检测时，应特别注意控制训练样本规模，如果数据量很大，支持向量机的训练时间就会比较长。

参考文献[5]提出了以域名的长度、域名的字符熵和元音字母所占比例等共 5 维特征，基于支持向量机构建分类器。本文认为，相对参考文献[4]，参考文献[5]使用的特征类型要丰富很多，在小样本训练集上的表现优异，模型的稳定性更高，但是模型的精度难以进一步提升，仍需要考虑不断扩充有效的特征。

2.3　基于深度学习的检测技术

深度学习是机器学习机制的复杂模型，源于人工神经网络的研究。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，具有将输入信息提取为最佳特征表示的巨大能力。与朴素贝叶斯、支持向量机等传统机器学习方法相比，深度学习可以避免专家定义的特征抽取步骤，能直接以无监督学习实现高级特征的抽取。在传统绝大多数的文本分类、语音识别和图像识别等应用场景中，深度学习方法都能取得比传统机器学习算法更好的分类效果。

人工智能技术被认为是新一代技术革命，受到普遍重视[6]，并逐步渗透到各个领域。参考文献[7]利用 LSTM 和 CNN 进行 DGA 域名检测，显著提升了对 DGA 域名的检测精度，但对于在训练样本集中较少出现的 DGA 域名家族表现不佳。此外，在同样的训练样本库条件中，深度学习的训练时间相对传统的机器学习算法较长，这在重视缩短应急响应时间的网络安全领域是一个明显的劣势。

3　基于组合分类器的恶意域名检测技术

在 DGA 域名检测中应用机器学习方法时，有如下两个难以回避的问题。

（1）训练样本天然不平衡

通常以全球 Alex Top*N* 的域名作为白名单，这是比较固定的样本集，而用于构建黑名单的 DGA 域名家族样本，主要依赖对僵尸网络程序的逆向破解 DGA 算法后，自动生成大量 DGA 域名。通常一个 DGA 域名家族的样本数以十万乃至上百万计。随着时间的积累，用于训练的白名单样本和黑名单样本“不平衡”的现象日益突出，将严重影响机器学习算法的应用效果。

（2）离线训练的时间越短越好

在满足一定准确率的前提下，算法的离线训练时间越短，意味着对新型网络威胁的检测时效性就越强，有利于提升整体的安全应急响应能力。

针对这些问题，本文设计了基于组合分类器的恶意域名检测技术，如图 2 所示，主要包括黑白名单过滤、Bigram/Trigram 特征抽取、朴素贝叶斯分类器模型、域名字符统计特征抽取和 SVM 分类器模型共 5 个关键模块。

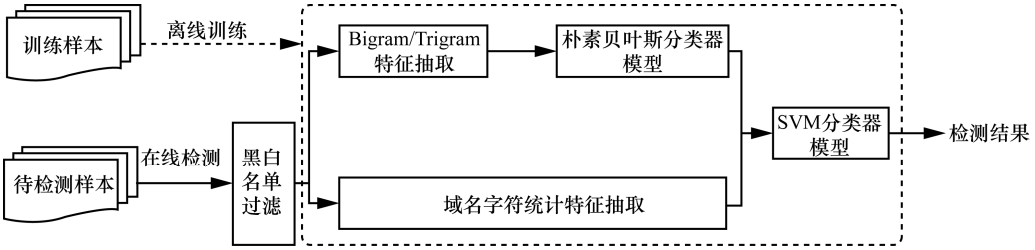


图 2　基于组合分类器的恶意域名检测框架

- 黑白名单过滤。在线监测过程中，通过设置一定规模的黑白名单库，包括 IP 地址黑/白名单库、域名黑/白名单库，可以对指定的 IP 地址或域名实现快速检测，满足网络安全的应急响应需求。在实际应用中，黑白名单库的规模需要严格控制，否则将严重影响系统的整体性能表现。
- Bigram/Trigram 特征抽取。在参考文献[4]中，已经证明了在 DGA 恶意域名检测中，Bigram/Trigram 特征的有效性。Bigram/ Trigram 特征将只用于朴素贝叶斯分类器模型。
- 朴素贝叶斯分类器模型。对于待检测样本 d，经过 Bigram/Trigram 特征抽取后，可以计算出 d 在各类别 C_i 的联合概率 $P(d, C_i)$。$P(d, C_i)$ 后续作为 SVM 分类器模型的输入特征。
- 域名字符统计特征抽取。包括域名长度、域名字符的元辅音比例等常用统计特征，作为 SVM 分类器模型的输入特征。
- SVM 分类器模型。SVM 分类器模型作为主分类器，在训练过程中，应特别注意控制训练样本库的平衡性，尽可能避免在两类训练样本中，其中某类样本个数远大于另一类的情况。

3.1　基于朴素贝叶斯分类器的特征抽取

全球 Alex Top$N$ 的域名列表比较稳定，很好地代表了具有较高经济价值、符合互联网用户的域名访问习惯的正常域名。例如，全球的 Alex Top 5 000 域名列表中，它们的 Bigram/Trigram 特征中具有较强的收敛性，一些特定的字符组合经常被复用。表 1 和表 2 为 Alex Top5 000/50 000/100 000 域名的 Bigram/Trigram 特征组合的统计情况。

表 1　Alex Top$N$ 域名的 Bigram 特征组合及其频率 Top10

Alex Top5 000	Alex Top50 000	Alex Top100 000
‘CO’：3 488	‘CO’: 31 963	‘CO’: 63 208
‘.C’：3 484	‘.C’: 31 802	‘.C’: 62 016
‘OM’：3 232	‘OM’: 29 387	‘OM’: 57 576
‘E.’,：641	‘IN’: 7 326	‘IN’: 15 535
‘IN’：628	‘NE’: 6 428	‘NE’: 13 260
‘NE’: 609	‘S.’: 6 408	‘S.’: 13 179
‘S.’: 571	‘OR’: 5 940	‘OR’: 12 685
‘AN’: 559	‘E.’: 5 920	‘ER’: 12 132
‘ER’: 499	‘ER’: 5 600	‘E.’: 11 902
‘OR’: 489	‘AN’: 5 449	‘AN’: 11 324

表 2　Alex Top$N$ 域名的 Trigram 特征组合及其频率 Top10

Alex Top5 000	Alex Top50 000	Alex Top100 000
‘.CO’: 3280	‘.CO’: 29 517	‘.CO’: 57 807
‘COM’: 3134	‘COM’: 28 210	‘COM’: 55 008
‘E.C’: 440	‘S.C’: 4 202	‘S.C’: 8 561
‘S.C’: 399	‘E.C’: 3 683	‘E.C’: 7 243
‘OM.’: 284	‘NET’: 3 010	‘NET’: 6 103
‘NET’: 280	‘.NE’: 2 506	‘.NE’: 5 048
‘A.C’: 244	‘OM.’: 2 474	‘OM.’: 4 643
‘R.C’: 239	‘T.C’: 2 273	‘T.C’: 4 555
‘T.C’: 238	‘R.C’: 2 076	‘A.C’: 4 081
‘.NE’: 226	‘A.C’: 2 054	‘.OR’:4 068

与 Alex Top*N* 的域名列表比较稳定相比，新出现的 DGA 域名难以预测。唯一确定的是，无论哪一个家族的 DGA 域名，与高价值的常见域名列表相比都具有明显的差异，其中可以体现在 Bigram/Trigram 字符组合。常见 DGA 家族域名样本构建了 5 万/10 万的样本库，其 Bigram/Trigram 特征统计情况见表 3 和表 4。

表 3　DGA 家族域名的 Bigram 特征组合及其频率 Top10

DGA 家族域名 5 万	DGA 家族域名 10 万
‘.C’: 12 677	‘OR’:14 454
‘CO’: 11 544	‘.C’: 14 066’
‘IN’: 10 134	IN’: 12 280
‘CC’: 9 154	‘.I’: 11 926
‘OM’: 9 090	‘RG’: 1 0857:
‘.I’: 9 052	‘RU’: 10 708
‘OR’: 6 509	‘BI’: 10 527
‘NF’: 6 395	‘NF’: 10 520
‘RG’: 6 265	‘FO’: 10 078
‘FO’: 6 257	‘CO’: 9 875

表 4　DGA 家族域名的 Trigram 特征组合及其频率 Top10

DGA 家族域名 5 万	DGA 家族域名 10 万
‘NET’: 12 412	‘ORG: 29 491
‘.NE’: 12 388	‘.OR’:29 395
‘ORG’: 12 029	‘NET’: 23 212
‘.OR’: 12 022	‘.NE’: 23 158
‘NS.’: 10 116	‘X.O’: 20 133
‘DDN’: 10 100	‘S.N’: 20 117
‘DNS’: 10 095	‘DNS’: 20 062
‘WW.’, 10 037	‘WWW’: 20 054
‘X.O’: 10 030	‘WW.’: 20 047
‘KHC’: 10 024	‘NS.’: 20 039

从表 1~表 4 可以看出，Bigram/Trigram 特征对于区分正常域名和 DGA 域名是有益的。因此，为了更有效地利用 Bigram/Trigram 的特征，同时与其他统计特征相融合，本文利用朴素贝叶斯分类器模型只训练了 Alex Top20 000 的域名，得到一个正常样本下的生成模型。对于待检测样本 *d*，依据此模型，计算得到 *P*(*d*, *C*)。如果待检测样本 *d* 为正常域名，则 *P*(*d*, *C*)较大；反之，如果 *d* 为 DGA 域名，则 *P*(*d*, *C* 正常)较小，为了更方便地进行对比，本文对概率值进行了取对数操作，Alex 域名和 DGA 域名的生成概率均值和方差对比如图 3 和图 4 所示。

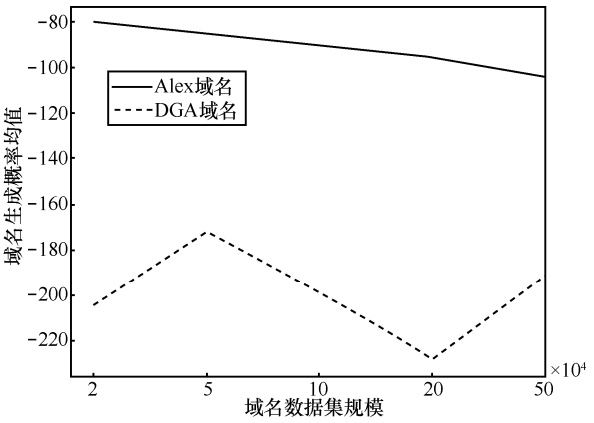


图 3　域名生成概率均值对比

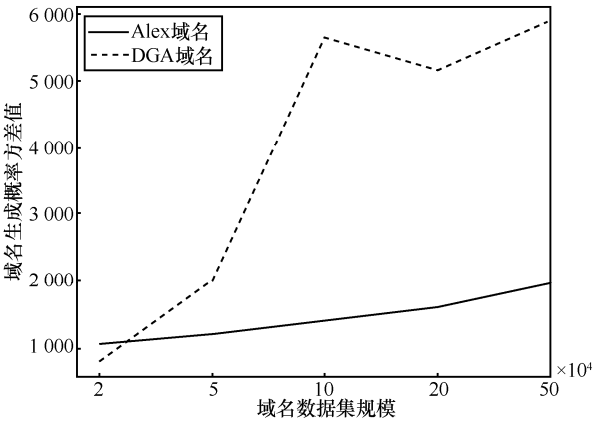


图 4　域名生成概率方差值对比

3.2　域名字符统计特征抽取

DGA 域名普遍具有以下特征：为了避免 DGA 域名和正常域名产生碰撞，相比正常域名 DGA 域名长度明显更长；DGA 域名基于输入的种子生成，具有一定随机性，域名中字母和数字的分布表现出无明显规律，这也是随机性算法的要求；

正常域名为了便于记忆和传播，尽可能讲究元辅音搭配，而 DGA 域名无此特征。作为 SVM 分类器模型的特征输入，本文还抽取 5 个域名的字符统计特征，具体包括以下几个方面。

（1）域名层级数

$$\text{特征1:}\frac{\text{域名层级数}}{\text{域名长度}}$$

攻击者为了避免和正常域名产生碰撞，不仅增加 DGA 域名长度，也会精心构造包含多个层级的恶意域名。为了用户有更好的体验，正常域名不仅仅是长度，域名的层级数也不会太多，因此域名层级数可以作为一个衡量 DGA 域名的特征。同时，为了能够使机器学习算法更快地收敛，为域名层级数除以域名字符长度，统一这一特征的量纲。

（2）元音占字符长度比

$$\text{特征2:}\frac{\text{元音字符数}}{\text{域名长度}}$$

众所周知，为了便于域名的记忆和传播，在构造域名时往往会选取可读性强的单词或者单词组合作为域名。根据语言学的规律，元音和辅音的组合明显更具有可读性。而在 26 个英文字母中，元音字母只有 a、e、i、o、u 5 个，剩余的全部是辅音字母。相反，由于 DGA 域名在生成的时候并不考虑元音和辅音的组合，因此 DGA 域名的元音占比会明显的比正常域名要低。

（3）字符集的离散度

$$\text{特征3:}\frac{\text{域名字符集合大小}}{\text{域名长度}}$$

基于 DGA 域名生成时的随机性，在英文字母和数字符中随机选取，会包含相当数目重复的字符，所以 DGA 域名中的包含字符集合大小除以域名长度后的数值远比正常域名的小。而正常域名长度虽然较短，而且为了易读性往往选取元音和辅音组合的字符，例如 weibo.com，baidu.com 等，所以正常域名的字符集合大小除以域名长度后会更高。

（4）域名长度

$$\text{特征4:}\frac{\text{域名长度}-\text{域名长度平均值}}{\text{域名长度标准差}}$$

为了有更好的用户体验，正常域名不会使用太长的域名，否则不便于域名的记忆和传播。而 DGA 域名恰巧相反，例如在僵尸网络中利用 DomainFlux 技术生成 DGA 域名作为 CC 服务器的通信域名，或者勒索病毒使用 DGA 域名作为恶意活动的开关，为了避免 DGA 域名和正常域名产生碰撞，阻断了恶意程序的执行和传播，攻击者必须使用长度更长的域名。所以域名长度是区分正常域名和恶意域名的一个重要特征。

而为了使得机器学习在计算中更好地收敛，本特征使用了归一化算法，使用正常域名的长度标准差和域名长度平均值对域名长度进行归一化。

（5）数字字母切换频率

$$\text{特征5:}\frac{\text{数字字母切换次数}}{\text{数字字母切换平均值}}$$

数字字母切换数目是指例如 a1b2.com 这样的域名，由“a”切换到“1”，再切换到“b”，再切换到“2”，最后切换到“.”，一共进行了 4 次数字和字母的切换。考虑 DGA 域名由字母和数字选取，并有一定的随机性，显然 DGA 域名的数字字母切换频率会更高。

4 实验与分析

4.1 数据集

本文的实验数据主要来自全球 Alex Top*N* 域名列表和 360 Netlab Open Data Project 的公开数据集。前者用于构建白样本训练库，后者用于构建黑样本训练库。进一步地，从 Alex Top*N* 中随机抽取了部分样本作为白样本测试集，从 CNCERT、360 Netlab Open Data Project 最新一年内公布的 DGA 域名随机抽取了部分样本作为黑样本测试集。为了更客观地衡量不同方法的效果，

本文构造了 5 个不同规模的数据集用于本次实验。数据集配置见表 5。

表 5　数据集配置情况

数据集编号	正负样本比例	训练集规模	测试集规模
1	1:1	10 000	2 000
2	1:1	20 000	4 000
3	1:1	40 000	8 000
4	1:1	60 000	12 000
5	1:1	100 000	20 000

4.2　代码运行环境

本实验主要是针对朴素贝叶斯分类器模型、普通 SVM 模型和 LSTM 模型，与本文的组合分类器进行比较，并对比各方面的评价指标，实验的软硬件开发环境见表 6。

表 6　实验软硬件开发环境

类别	环境设置
操作系统	Ubuntu18.0.4 LTS
开发语言	Python3.6
开发工具	PyCharm 企业版 v2019.1
机器学习库	Sklearn, TensorFlow
CPU	Intel Core i7-6700HQ @2.60 GHz
内存	16 GB
硬盘	1 TB

4.3　结果分析

其中，朴素贝叶斯分类器模型采用了 Bigram 和 Trigram 相结合的特征抽取方法；普通 SVM 模型采用了本文所定义的 5 个字符统计特征；LSTM 模型不需要特征抽取。为了从离线训练时间、检测效果多个维度对模型进行评估，本文采用的评价指标为训练时长、正确率、精准率、召回率、*F*1 值 5 个指标，在 5 个不同规模的数据集上进行比较。本文采用的组合分类器，包含朴素贝叶斯分类器和 SVM 模型，由于朴素贝叶斯分类器模型的训练集固定，在应用中只对 Alex Top20 000 的域名进行训练，得到模型后，直接应用于联合概率计算，故不计入组合分类器的训练时间。

本文根据混淆矩阵，计算各模型的正确率、精准率、召回率和 *F*1 值见表 7。混淆矩阵中的 4 个变量定义如下：

（1）真正（true positive，TP），被模型预测为正的正样本；

（2）假正（false positive，FP），被模型预测为正的负样本；

（3）假负（false negative，FN），被模型预测为负的正样本；

（4）真负（true negative，TN），被模型预测为负的负样本。

表 7　混淆矩阵

	预测为正例	预测为负例
真实类别为正例	真正(TP)	假负（FN）
真实类别为负例	假正(FP)	真负（TN）

而正确率、精准率、召回率和 *F*1 值的定义如下。

（1）正确率（accuracy）用于描述预测正确的样本占总体样本的比例，计算式如下：

$$\text{accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{1}$$

（2）精准率（precision）用于描述预测为正例的样本中有多少确实为正例，计算式如下：

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

（3）召回率（recall）用于描述真实类型为正例的样本有多少被预测出来，计算式如下：

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

（4）*F*1 值主要是为了融合精准率与召回率，做一个整体的评价，计算式如下：

$$F1=\frac{2\cdot\text{pecision}\cdot\text{recall}}{\text{pecision}+\text{recall}} \tag{4}$$

实验结果对比如图 5~图 9 所示。

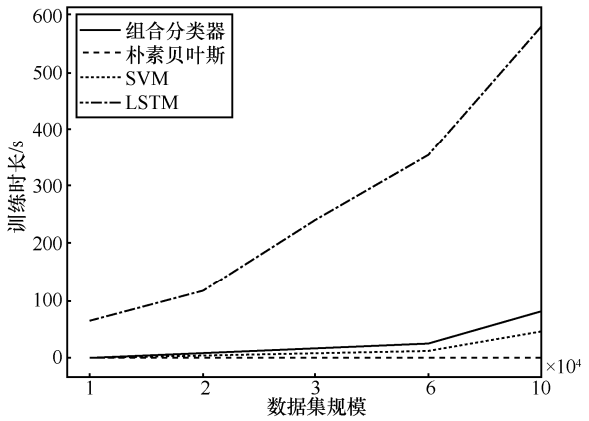


图 5　训练时长对比

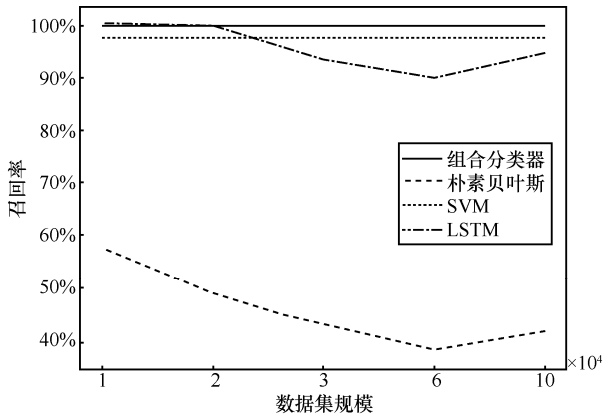


图 8　召回率对比

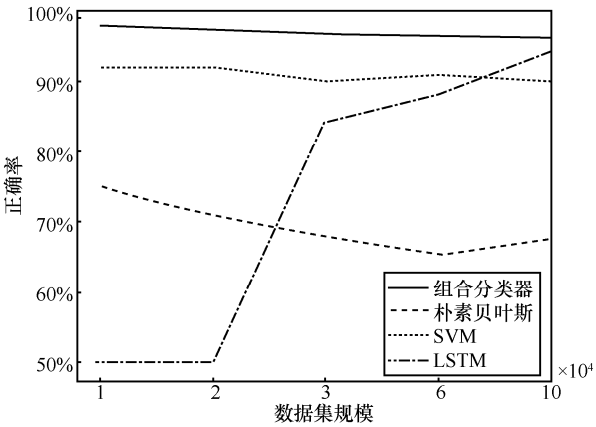


图 6　正确率对比

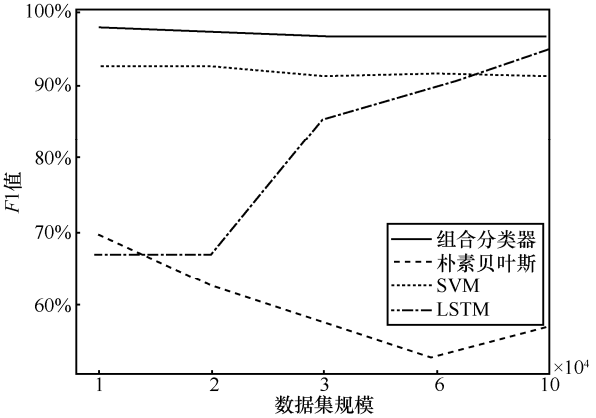


图 9　*F*1 值对比

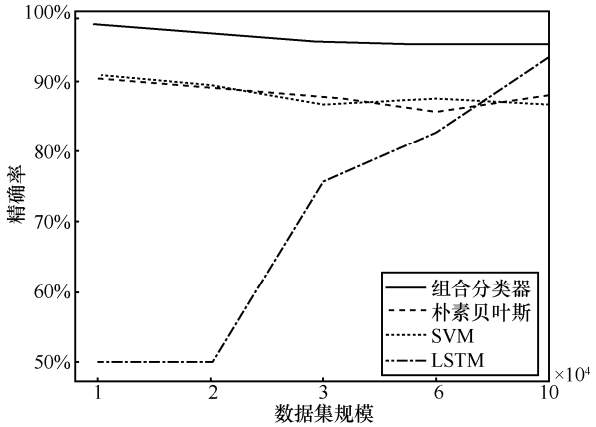


图 7　精准率对比

根据图 5~图 9 分析，组合分类器相对于朴素贝叶斯模型、SVM 模型和 LSTM 模型都具有明显的优势。本文的组合分类器模型的训练时长比朴素贝叶斯模型稍多一些，和普通的 SVM 模型相当，远好于 LSTM 模型。从检测效果来看，本文的组合分类器模型，正因为融合了朴素贝叶斯特征，使得组合器模型训练时的收敛速度比普通的 SVM 模型要快，在训练样本集不多的情况下，组合分类器仍然能达到相当不错的性能。此外，随着训练样本的数量增加，组合分类器的正确率、召回率、正确率和 *F*1 值一直都比较稳定。

5　结束语

随着技术生态的成熟，云计算业务发展迅猛，公有云平台因价格便宜、使用便捷和带宽可控，愈发受到网络犯罪分子的青睐。根据参考文献[8]统计，2019 年超过 90%的 CC 位于云或 VPS 服务器上，相比 2018 年的 79%有了进一步提升。IT 基础设施与实际业务主体的分离，作为域名和 IP 地址的映射解析协议，DNS 的重要性不断上升。一方面，基于 DNS 协议机制的各类攻击手段不断进步，如 Fast-Flux、Domain-Flux 技术在各类大型

僵尸网络中日趋普遍，基于 DNS 解析日志的恶意域名检测技术得到了业界的高度重视；另一方面，云计算业务的迅猛发展，也催生了 CDN、云 WAF 等业务的繁荣，而后者普遍使用 DGA 技术快速生成域名，这给基于机器学习的恶意域名检测技术带来极大的困扰，会导致极高的误报率。在实际的应用中，可通过设置二级域名白名单、IP 地址段白名单等规则，进行过滤，降低误报。长远来看，由于恶意域名的时效性较短，对 Fast-Flux、Domain-Flux 等恶意域名的检测，进一步挖掘相关的恶意域名家族，可以产生更高的威胁情报价值，恶意域名检测分析的相关工作仍有很大的探索空间。

参考文献：

[1] 安天安全研究与应急处理中心. 安天针对勒索蠕虫“魔窟”(WANNACRY) 的深度分析报告[R]. 2017.
Antiy CERT. An in-depth analysis report on the blackmail worm WANNACRY[R]. 2017.

[2] SkyEye 天眼实验室. APT OceanLotus 数字海洋的游猎. 2015.
SkyEye Laboratory. APT OceanLotus digital ocean safari[EB]. 2015.

[3] 臧小东, 龚俭, 胡晓艳. 基于 AGD 的恶意域名检测[J]. 通信学报, 2018, 39(7): 15-25.
ZANG X D, GONG J, HU X Y. Detecting malicious domain names based on AGD[J]. Journal on Communications, 2018, 39(7): 15-25.

[4] 蒋鸿玲, 戴俊伟. DGA 恶意域名检测方法[J]. 北京信息科技大学学报(自然科学版), 2019, 34(5): 45-50.
JIANG H L, DAI J W. DGA malicious domain name detection method[J]. Journal of Beijing Information Science & Technology University, 2019, 34(5): 45-50.

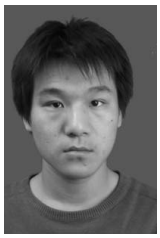
[5] 王震. 基于 SVM 的 DGA 域名检测方法研究[D]. 山东: 济南大学, 2018.
WANG Z. Research on DGA domain name detection method based on SVM[D]. Shandong: Jinan University, 2018.

[6] 王志宏, 杨震. 人工智能技术的哲学及系统性思考[J]. 电信科学, 2018, 34(4): 12-21.
WANG Z H, YANG Z. Philosophy and systematic thinking of artificial intelligence technology[J]. Telecommunications Science, 2018, 34(4): 12-21.

[7] YU B, GRAY D L, PAN J, et al. Inline DGA Detection with Deep Networks[C]//Proceedings of IEEE International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2017: 683-692.

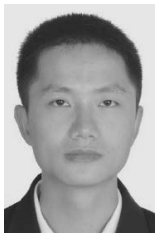
[8] 绿盟科技. 2019 年 BOTNET 趋势报告[R]. 2020.
NSFOCUS. BOTNET trends report 2019[R]. 2020.

[作者简介]



盛剑涛（1994– ），男，中国电信股份有限公司研究院工程师，主要研究方向为网络安全、机器学习。

陈茂飞（1986– ），男，中国电信股份有限公司研究院工程师，主要研究方向为网络安全、恶意程序分析。



刘东鑫（1985– ），男，中国电信股份有限公司研究院工程师，主要研究方向为网络安全、大数据安全。

汪来富（1976– ），男，中国电信股份有限公司研究院高级工程师，主要研究方向为大数据安全、云计算安全、网络安全。

史国水（1972– ），男，中国电信股份有限公司研究院工程师，主要研究方向为网络安全。

金华敏（1972– ），男，中国电信股份有限公司研究院高级工程师，主要研究方向为 IP 网、云计算、大数据安全以及网络安全。