基于XGBoost和粒子群优化算法的DGA恶意域名识别

陈泽生，周敏，冯李春，陈伟杰
（广州美术学院信息技术中心，广东 广州 510006）

摘　要：恶意域名生成算法（DGA）已成为一种常见的网络攻击手段，为了提高对DGA恶意域名的检测能力，提出了一种基于XGBoost和粒子群优化（PSO）算法的恶意域名识别方法。首先，以交叉验证准确率作为评估指标，使用PSO算法对XGBoost进行超参数寻优，然后基于XGBoost进行分类识别。实验结果显示，经过PSO优化的XGBoost模型在DGA恶意域名分类识别中性能得到提升，相较于其他分类模型，在准确率、精确率、召回率和F1分数等评价指标上获得了更优的效果。研究表明，结合PSO算法进行参数能够有效地提升XGBoost模型在DGA恶意域名识别任务中的表现。

关键词：域名生成算法；XGBoost；粒子群优化；特征选择
中图分类号：TP3.2.2
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024237

DGA malicious domain name identification based on XGBoost and particle swarm optimization algorithm

CHEN Zesheng, ZHOU Min, FENG Lichun, CHEN Weijie
Information Technology Center, Guangzhou Academy of Fine Arts, Guangzhou 510006, China

Abstract: Domain generation algorithms (DGA) have become a common method of network attacks. To enhance the detection capability for DGA malicious domains, a method for malicious domain identification based on XGBoost and particle swarm optimization (PSO) algorithms was proposed. Firstly, using cross-validation accuracy as the evaluation metric, the PSO algorithm was employed to optimize the hyperparameters of XGBoost, followed by classification and identification using XGBoost. Experimental results demonstrate that the XGBoost model optimized by PSO exhibits improved performance in DGA malicious domain classification. Compared to other classification models, it achieves better results in metrics such as accuracy, precision, recall, and F1_score. The study indicates that integrating PSO for parameter selection effectively enhances the performance of XGBoost in DGA malicious domain identification tasks.

Keywords: domain generation algorithm, XGBoost, particle swarm optimization, feature selection

0　引言

目前互联网已成为现代社会的基础设施，支撑着全球的信息交流、商业活动和社会运作。然而，互联网的普及和开放性也带来了诸多安全挑战。网络攻击活动日益猖獗，在国家互联网应急中心[1]发布的近几年中国互联网安全报告和安全态势中，可以发现各种恶意程序、安全漏洞等面临着严峻的形势。其中，僵尸网络（Botnet）作为一种自动化、分布式的网络攻击手段，对网络安全构成了严重威胁[2]。僵尸网络通过感染大量计算机或其他网络设备，形成庞大的控制网络，攻击者可以远程操控这

收稿日期：2024–10–21
通信作者：周敏，zhoumin@gzarts.edu.cn
基金项目：广州美术学院学术提升计划基金资助项目（No.24XS38）
Foundation Item: Guangzhou Academy of Fine Arts Academic Improvement Program（No.24XS38）

些设备进行恶意活动，如分布式拒绝服务（DDoS, distributed denial of service）攻击、垃圾邮件、挖矿、勒索、信息窃取等。僵尸网络的核心是其命令与控制（C&C, command and control）服务器，它负责接收攻击指令并分发至网络中的各个节点。为了隐藏C&C服务器的真实身份和位置，避免被安全检测系统发现，攻击者采用了一种名为域名生成算法（DGA, domain generation algorithm）的技术。DGA能够自动生成大量看似随机的域名，作为C&C服务器的代理域名，使僵尸网络的控制和通信更加隐蔽和难以追踪[3]。僵尸网络示意如图1所示。

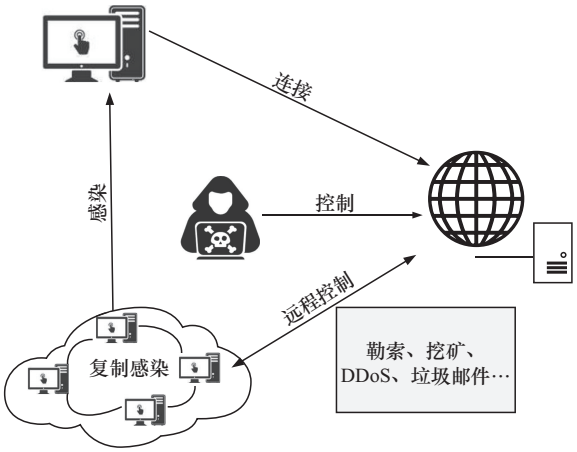


图1　僵尸网络示意

DGA域名与传统的域名注册方式存在显著差异，它们通常不包含有意义的词汇或易于记忆的模式，这为安全研究人员和自动化系统识别和拦截恶意域名带来了巨大挑战。随着DGA技术的发展，僵尸网络的存活周期得以延长，攻击能力不断增强，给网络安全防护带来了前所未有的压力。因此，研究和开发新的DGA恶意域名检测技术，对于提高网络安全防护能力具有重要意义。

目前，机器学习模型的性能很大程度上依赖于其超参数的设置。超参数优化是一个复杂且计算成本高昂的过程，尤其是在面对大规模数据集时。为了解决这一问题，研究者们提出了多种超参数优化方法，包括网格搜索、随机搜索以及基于启发式的优化算法等。粒子群优化（PSO, particle swarm optimization）算法作为一种模拟鸟群或鱼群行为的优化算法，因其简单、高效和易于实现的特点，在解决优化问题上显示出了巨大的潜力[4]。

XGBoost算法作为梯度提升决策树（GBDT, gradient boosting decision tree）的一种实现，因其出色的预测性能和计算效率，在各种数据挖掘和机器学习竞赛中取得了优异的成绩。XGBoost通过正则化项和梯度下降方法，有效地解决了传统GBDT模型的过拟合问题，并提高了模型的泛化能力。但是，XGBoost模型的超参数选择同样对模型性能有着决定性的影响。

为了提高DGA恶意域名的检测能力，本文提出了一种基于XGBoost和PSO算法的恶意域名识别方法。该方法首先利用PSO算法对XGBoost模型的超参数进行优化，以寻找最佳的参数组合；然后使用优化后的XGBoost模型进行恶意域名的分类识别。

1　相关工作

在DGA恶意域名检测领域，国内外众多研究者开展了大量工作，提出了多种检测方法，早期较为简单的黑名单方法难以应对大量DGA域名[5]。目前的DGA检测包括基于统计分析、机器学习、深度学习的各种DGA识别技术。从DGA本身可以大概划分为基于DGA字符特征、基于DGA关联特征、基于DGA域名本身三大类；从技术角度，目前的主流方法又可以划分为基于传统机器学习检测方法、基于无监督学习的检测方法、基于深度学习的检测方法。

例如，黄凯等[6]提出一种基于字符及解析特征的恶意域名检测方法，设计了域名的字符统计特征、相似度特征、解析特征等，并利用C4.5决策树进行分类识别。Vranken等[7]使用TF-IDF来测量域名中最相关的n元语法的频率，并将其用作学习算法的特征。赵正利等[8]利用粒子群算法进行特征选择，然后使用SVM进行恶意域名检测。Hoang等[9]使用随机森林进行DGA检测，达到97.03%的准确率。

随着深度学习的发展，得益于神经网络自动从域名字符串中学习高级特征，而不需要人工手动提取特征，基于深度学习的DGA检测也取得了很多的研究成果。例如，盛振威等[10]提出一种融合卷积神经网络和门控循环单元网络的DGA检测深度学习模型，有效提取域名信息里隐藏的局部特征和上下文关联性特征。Shahzad等[11]提出一种RNN架

构对多个域名数据集进行评估。

近年来，注意力机制在多个领域展现出其强大的能力，特别是在处理序列数据时。在DGA域名检测中，林思明等[12]提出一种基于BiLSTM神经网络的DGA域名检测方法，采用词袋模型对域名进行处理，将字符类型的域名转换为适合BiLSTM神经网络的输入数据，然后基于BiLSTM神经网络设计适合DGA域名检测的各层神经网络的参数，从而实现DGA域名的检测。闫莉莉[13]采用多个并行的CNN模型，并引入了带有注意力机制的BiLSTM模型。郝旭光[14]通过结合输入层、Embedding层、卷积神经网络层、注意力模块和长短时记忆网络层，实现层次化特征提取使模型性能得到较大改善。

2　模型算法原理

XGBoost是一种通过集成学习的方法将多个弱分类器组合成一个强分类器。其基本思想是给定一个训练集$\{(x_i y_i)\}_{i=1}^{n}$，其中，x_i是输入特征，y_i是目标变量，先初始化预测为常数值，然后在每一步中建立一个新的模型来拟合当前模型的损失函数，最终的预测值是所有模型预测结果的加权和。其目标函数如式(1)所示。

$$\mathrm{Obj}(t)=\sum_{i=1}^{n}L\left(y_i\hat{y}_i^{t-1}+f_t(x_i)\right)+\Omega(f_t) \quad (1)$$

其中，$L\left(y_i\hat{y}_i^{t-1}+f_t(x_i)\right)$表示在第$t$轮，新的模型$f_t(x)$对目标变量$y_i$的损失；$\Omega(f_t)$表示对模型复杂度的惩罚。

PSO算法是一种群体智能的优化算法，通过模拟个体之间的信息共享与协作，来寻找问题的最优解。其基本思想是，在搜索空间中模拟一群“粒子”的运动，每个粒子代表一个潜在解，粒子通过跟踪自身的历史最佳位置和全局最佳位置来更新其速度和位置，从而逐步逼近最优解。每个粒子的位置向量$\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})$表示问题的一个潜在解，其中，$i$是粒子的索引，$d$问题的维度。而每个粒子的速度向量$\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{id})$表示粒子在各个维度上的变化速率。PSO的算法步骤可以简化如下。

初始化

1) 设定粒子数目N，最大迭代次数T

2) 随机初始化粒子的速度$\boldsymbol{V}$和位置$\boldsymbol{X}$

3) 设定每个粒子的最佳位置pbest为其当前位置$\boldsymbol{X}$,设定全局最佳位置gbest为所有粒子的pbest中的最佳位置

主循环

1) 对于每个粒子i:

① 计算粒子i的适应度值（目标函数值）

② 如果当前适应度值优于粒子i的pbest，则更新pbest为当前粒子位置

③ 如果粒子i的pbest优于全局最佳位置gbest，则更新gbest为pbest

2) 更新每个粒子的速度和位置

① 对于每个粒子i计算速度更新公式：$\boldsymbol{V}_i^{t+1}=wv_i^t+c_1r_1(\mathrm{pbest}_i^t-\boldsymbol{X}_i^t)+c_2r_2(\mathrm{gbest}^t-\boldsymbol{X}_i^t)^t$，其中，$w$为惯性权重，$c_1$和$c_2$为学习因子，$r_1$和$r_2$为介于0和1之间的随机数，$t$表示当前的迭代

② 对于每个粒子i更新位置$\boldsymbol{X}_i^{t+1}=\boldsymbol{X}_i^t+\boldsymbol{V}_i^{t+1}$

检查终止条件

1) 如果达到最大迭代次数T，则终止算法

2) 否则，返回到主循环

输出

1) 最佳位置gbest

2) 对应的适应度值

3　实验结果与分析

3.1　实验环境

本文实验在WINDOWS10系统上进行，硬件环境处理器为Intel(R) Core(TM) i5-4200H CPU @ 2.80 GHz，运行内存为16 GB, GPU计算资源为NVIDA GeForce GTX 950M, 算法实现基于Python, Scikit-learn机器学习库进行实现。

3.2　数据集及特征抽取

本文所采用的数据集主要分为两部分，一部分是DGA域名数据集，来自著名的DGArchive数据集，另一部分来自Cisco umbrella基于被动DNS情况统计的前100万热门域名。本文选取了DGArchive 2020-06-19发布的数据集，剔除样本量小于2 000的DGA后，从剩下的每个类型的DGA数据集中选取前2 000个样本，组成实验DGA数据集。共61类DGA家族，每个家族2 000个样本，共122 000个样本，基本情况如表1所示。此外，本文从Cisco umbrella发布的数据集中，选取与DGA数量相等的正常域名，该数据集仅有两列特征，分别是序号和域名。本文将这两类数据集合并组成完

表1 DGA家族基本情况

编号	DGA家族	编号	DGA家族	编号	DGA家族
1	bamital_dga	21	murofetweekly_dga	41	ranbyus_dga
2	banjori_dga	22	murofet_dga	42	rovnix_dga
3	bedep_dga	23	mydoom_dga	43	shifu_dga
4	blackhole_dga	24	necurs_dga	44	simda_dga
5	chinad_dga	25	nymaim2_dga	45	sisron_dga
6	conficker_dga	26	nymaim_dga	46	sphinx_dga
7	corebot_dga	27	oderoor_dga	47	suppobox_dga
8	cryptolocker_dga	28	padcrypt_dga	48	sutra_dga
9	diamondfox_dga	29	pandabanker_dga	49	symmi_dga
10	dnschanger_dga	30	pitou_dga	50	szribi_dga
11	dyre_dga	31	proslikefan_dga	51	tinba_dga
12	ekforward_dga	32	pushdotid_dga	52	tinynuke_dga
13	emotet_dga	33	pushdo_dga	53	tofsee_dga
14	gameover_dga	34	pykspa2s_dga	54	torpig_dga
15	gameover_p2p	35	pykspa_dga	55	ud2_dga
16	gozi_dga	36	qadars_dga	56	urlzone_dga
17	infy_dga	37	qakbot_dga	57	vawtrak_dga
18	locky_dga	38	qsnatch_dga	58	vidro_dga
19	matsnu_dga	39	ramdo_dga	59	virut_dga
20	moneromine r_dga	40	ramnit_dga	60	wd_dga
				61	xxhex_dga

整的实验数据集，域名样本量共244 000条。表2是部分DGA域名和正常域名示例。

表2 部分DGA域名和正常域名示例

DGA域名	正常域名
47faeb4f1b75a48499ba14e9b1cd895a.org	google.com
sgjprsensinaix.com	www.amazon.com
cybxcikisigkmqtl.ru	baidu.com
vinqsbp.net	data.microsoft.com

为了高效率地进行DGA恶意域名识别，进行特征抽取是非常有必要的。因此，本文基于域名字符串进行特征提取如下。

域名长度为

$$L = \mathrm{len}(\mathrm{domain}) \tag{2}$$

其中，domain表示域名字符串，len()表示取字符串长度。

元音字母比例为

$$V = \frac{N_v}{L} \tag{3}$$

其中，N_v表示元音字符数量。

数字字符比例为

$$D = \frac{N_d}{L} \tag{4}$$

其中，N_d表示数字字符数量。

连续字母比例为

$$C = \frac{N_c}{L} \tag{5}$$

其中，N_c连续字母对的数量。

连续数字比例为

$$C_d = \frac{N_{cd}}{L} \tag{6}$$

其中，N_{cd}表示连续数字对的数量。

不同字符数为

$$U = \mathrm{len}(\mathrm{set}(\mathrm{domain})) \tag{7}$$

其中，set(domain)表示字符串中不同字符的集合。

二级域名长度为

$$L_s = \text{len}(\text{sec ond\_level\_domain}) \tag{8}$$

域名熵值为

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{9}$$

其中，p_i是字符i的出现频率。

元音字符个数为

$$N_v = \sum_{i \in \text{vowels}} \text{count}(i) \tag{10}$$

其中，vowels是元音字符集合，$\text{count}(i)$是字符i的数量。

辅音字符个数为

$$N_c = \sum_{i \in \text{consonants}} \text{count}(i) \tag{11}$$

其中，consonants表示辅音字符集合。

特殊字符个数为

$$N_s = \sum_{i \in \text{species}} \text{count}(i) \tag{12}$$

其中，species表示特殊字符集合。

数字字符个数为

$$N_d = \sum_{i \in \text{digits}} \text{count}(i) \tag{13}$$

其中，digits表示数字字符集合。

字符的平均ASCII值为

$$A = \frac{\sum_{i=1}^{L} \text{ord}(\text{domain}[i])}{L} \tag{14}$$

其中，$\text{ord}(\text{domain}[i])$第$i$个字符的ASCII值。

3.3　算法验证分析

采用上述数据集并经过特征提取，特征归一化等数据处理之后，基于上述算法模型进行实验验证。实验采取训练数据集与测试数据集7:3的比例，选取准确率（Accuracy）、精确率（Precision）、召回率（Recall）作为评价指标，计算方式分别如下

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$F_{1\_score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

本文PSO设置粒子群swarm数量为100，迭代次数maxiter为50，对XGBoost进行超参数寻优，learning_rate设置取值范围是[0.01,1]，max_depth设置取值范围是[3,30]，subsample设置取值范围是[0.5,1.0]，训练过程中采用5折交叉验证，最终超参数设定learning_rate为0.21，max_depth为10，subsample为1.0。最终获得的性能指标如表3所示。

表 3　实验性能指标

性能指标	值
Accuracy	94.35%
Precision	96.54%
Recall	92.03%
F1_score	94.23%

基于PSO优化后的XGBoost混淆矩阵如图2所示。对于DGA域名，XGBoost实现了96.69%预测正确，3.31%的DGA被预测为正常域名。对于正常域名，XGBoost有92.03%识别为正常域名，7.97%识别为DGA域名。ROC曲线如图3所示。

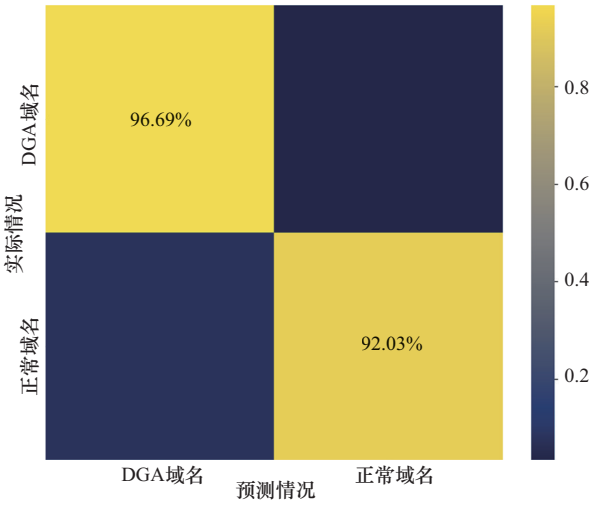


图 2　基于PSO优化后的XGBoost混淆矩阵

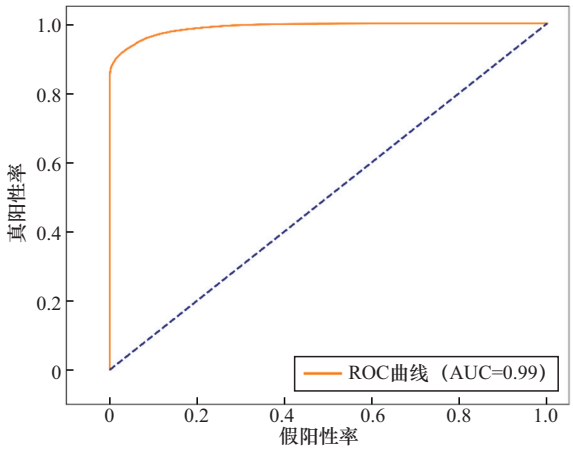


图 3　ROC曲线

为了对比算法的有效性，本文采用逻辑回归、最近邻、决策树、支持向量机对同样的数据集进行分类识别。评价指标对比结果如表4所示。

表4 各模型评价指标对比

算法	Accuracy	Precision	Recall	F1_score
逻辑回归	91.34%	94.50%	87.82%	91.03%
最近邻	93.41%	94.97%	91.69%	93.30%
决策树	93.03%	93.96%	92.00	92.97
支持向量机	92.67%	98.13%	86.95%	92.19%
XGBoost	94.35%	96.54%	92.03%	94.23%

4 结束语

本文基于XGBoost和粒子群优化算法对DGA恶意域名进行识别，手工提取常见的域名特征后经数据预处理和特征归一化形成数据集。使用PSO对XGBoost进行超参数寻优。使用XGBoost进行数据验证，验证过程中在训练集上进行交叉验证，避免训练过拟合，然后在测试集上进行分类。在多达61类DGA家族组成的恶意域名中，获得了整体94.35%的准确率，而识别DGA恶意域名的准确率达到96.69%。与其他机器学习算法相比，在Accuracy、Recall、F1_score中均获得更优的性能。

参考文献：

[1] 国家互联网应急中心. 2020年中国互联网网络安全报告[R]. (2021-07-21)[2024-08-11].

[2] 俞意, 李建华, 沈晨, 等. IoT僵尸网络传播大规模测量研究[J]. 计算机时代, 2023(9): 37-42, 47.
YU Y, LI J H, SHEN C, et al. Large-scale measurement study of IoT botnet infection behavior[J]. Computer Era, 2023(9): 37-42, 47.

[3] 赵科军. 基于深度学习的DGA域名检测与生成方法研究[D]. 济南: 山东大学, 2024.

[4] 杨帆,乌景秀,范子武,等. 快速综合学习粒子群优化算法[J/OL]. 水利水电技术(中英文), (2024-07-24)[2024-10-20].

[5] KÜHRER M, ROSSOW C, HOLZ T. Paint it black: evaluating the effectiveness of malware blacklists[C]//Proceedings of Research in Attacks, Intrusions and Defenses. Cham: Springer International Publishing, 2014: 1-21.

[6] 黄凯, 傅建明, 黄坚伟, 等. 一种基于字符及解析特征的恶意域名检测方法[J]. 计算机仿真, 2018, 35(3): 287-292.
HUANG K, FU J M, HUANG J W, et al. A malicious domain detection approach based on character and resolution features[J]. Computer Simulation, 2018, 35(3): 287-292.

[7] VRANKEN H, ALIZADEH H. Detection of DGA-generated domain names with TF-IDF[J]. Electronics, 2022, 11(3): 414.

[8] 赵正利, 姜鹏, 仲国强, 等. 基于SVM-RFE和粒子群优化算法的恶意域名检测模型[J]. 福州大学学报(自然科学版), 2023, 51(5): 634-638.
ZHAO Z L, JIANG P, ZHONG G Q, et al. A SVM-RFE and particle swarm optimization based detection model for malicious domain names [J]. Journal of Fuzhou University (Natural Science Edition), 2023, 51 (5): 634-638.

[9] HOANG X D, VU X H. An improved model for detecting DGA botnets using random forest algorithm[J]. Information Security Journal: A Global Perspective, 2022, 31(4): 441-450.

[10] 盛振威, 徐国天. 基于融合CNN与GRU的DGA恶意域名检测方法[J]. 网络安全技术与应用, 2022(12): 29-32.
SHENG Z W, XU G T. Detection method of DGA malicious domain Name based on fusion of CNN and GRU[J]. Network Security Technology & Application, 2022(12): 29-32.

[11] SHAHZAD H, SATTAR A R, SKANDARANIYAM J. DGA domain detection using deep learning[C]//Proceedings of 2021 IEEE 5th International Conference on Cryptography, Security and Privacy. Piscataway: IEEE Press, 2021: 139-143.

[12] 林思明, 陈腾跃, 梁煜麓. 基于BiLSTM神经网络的DGA域名检测方法[J]. 网络安全技术与应用, 2019(1): 15-17.
LIN S M, CHEN T Y, LIANG Y L. Detection method of DGA domain Name based on BiLSTM neural network[J]. Network Security Technology & Application, 2019(1): 15-17.

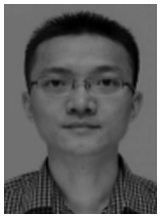
[13] 闫莉莉. 基于神经网络的恶意DGA域名检测技术研究[D]. 济南: 齐鲁工业大学, 2024.
YAN L L. Research on malicious DGA domain Name detection technology based on neural network[D]. Jinan: Qilu University of Technology, 2024.

[14] 郝旭光. 基于注意力特征融合网络的DGA恶意域名检测方法[J]. 网络安全与数据治理, 2024, 43(1): 19-27.
HAO X G. A DGA malicious domain detection method based on attention feature fusion network[J]. Cyber Security and Data Governance, 2024, 43(1): 19-27.

[作者简介]

陈泽生（1979– ），男，广东汕头人，广州美术学院正高级工程师，主要研究方向为计算机网络、网络信息安全、计算机应用等。

周敏（1993– ），男，江西上饶人，广州美术学院高级工程师，主要研究方向为网络空间安全、信息化建设等。



冯李春（1985– ），男，广东湛江人，广州美术学院工程师，主要研究方向为网络空间安全、云计算技术。

陈伟杰（1996– ），男，湖南郴州人，广州美术学院工程师，主要研究方向为信息化建设、大数据分析等。