

基于改进的 SVM 方法的异常检测研究

张辉¹, 刘成²

(1. 新疆公安厅特别侦察队, 新疆 乌鲁木齐 830000;
2. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘要: 利用非参数检验的方法提取出对分类结果影响显著的特征变量, 提出一种改进的 SVM 多分类方法 (D-SVM), 其融合了判别分析, 可以解决样本不均衡导致的分类不准确和误报率高的问题。将多分类问题处理成一个个二分类问题, D-SVM 既可以保持 SVM 较好的分类准确性, 同时又可以不受样本不均衡的影响, 具有较低的误报率。将 D-SVM 应用到 KDD99 数据集, 结果表明, 该方法具有较高的分类准确性和较低的误报率。

关键词: 异常检测; 非参数检验; SVM 分类; 样本不均衡; 判别分析

中图分类号: TP309/TP274

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00092

Anomaly intrusion detection based on modified SVM

ZHANG Hui¹, LIU Cheng²

(1. Special Reconnaissance Team of Xinjiang Public Security Bureau, Urumpi 830000, China;
2. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

Abstract: A modified SVM multi-classification algorithm integrated with discriminant analysis (D-SVM) was proposed, which could solve the problem of low detection accuracy and high false alarm rate caused by unbalanced datasets. For a multi-classification problem could be divided into several binary classification problems, D-SVM could not only have the virtue of high detection accuracy, but also have a low false alarm rate even confronted with unbalanced datasets. Experiments based on KDD99 dataset verify the feasibility and validity of the integrated approach. Results show that when confronted with multi-classification problems, D-SVM could achieve a high detection accuracy and low false alarm rate even when SVM alone fails because of the unbalanced datasets.

Key words: anomaly detection, non-parametric test, SVM classifier, unbalanced datasets, discriminant analysis

1 引言

计算机技术和互联网的发展极大地促进了网络与企业、个人生活的融合。据中国互联网信息中心《第37次中国互联网络发展状况统计报告》显示,截至2015年12月,全国使用计算机办公的企业比例为95.2%,涉及供应链、营销、财务和人力资源等方面;与此同时,网络购物和网络社交等已成为人们生活必不可少的一部分。随着

“互联网+”的逐渐深入,网络的作用会更加突出,网络安全问题也将会变得十分严峻,并成为制约网络发展的因素。入侵检测在网络安全中有重要的作用,它可以实时地监测、阻止来自网络外部和内部的入侵,保护网络免受攻击,造成损失。入侵检测方法可以分为2类^[1]:误用检测(misuse detection)和异常检测(anomaly detection)。误用检测只能检测已知的攻击,而异常检测却可以检测新的、未知的攻击。异常检测已成

收稿日期: 2016-06-11; 修回日期: 2016-07-23。通信作者: 刘成, lc@cert.org.cn

为入侵检测领域的主要研究对象^[2,3], 因此, 本文以异常检测为主要研究内容。

异常检测的实质是分类问题, 即如何将数据分为正常和异常 2 类。相关研究领域已有很多研究成果。异常检测的研究方法主要有基于统计的方法、基于数据挖掘的方法和基于机器学习的方法这几类^[4]。基于统计的方法其核心内容就是利用统计方法设定阈值^[5,6]或概率^[7,8], 对于未知的连接或数据分组, 检验其是否在设定的阈值或概率范围内, 从而判定是否为入侵或攻击。基于数据挖掘的方法是以数据为中心, 利用数据挖掘的相关技术和算法, 找出审计数据或流量数据中存在的规律, 从而发现入侵行为^[9]。基于数据挖掘的检测方法主要有基于离群点的挖掘方法、基于分类的检测方法、基于聚类的检测方法和基于关联分析的检测方法^[10,11]。基于机器学习的方法主要有神经网络、遗传算法、隐马尔可夫模型和支持向量机 (SVM) 等^[12]。

在这些方法中, SVM 具有比较好的检测效果^[13,14]。SVM 是在统计学习理论的基础上, 充分考虑了结构风险最小化, 通过学习得到一个使分类间隔最大化的超平面, 从而将不同的类分别开来。同时, SVM 具有良好的泛化能力, 已经被广泛应用到人脸识别、数据挖掘、入侵检测等领域, 并取得良好的分类效果。本文选取 SVM 作为入侵检测方法, 可以达到比较好的检测效果。但 SVM 方法对训练样本有一定的要求, 如果训练不足, 可能会导致 SVM 分类效果和进度达不到要求。因此, 本文在 SVM 方法的基础上引进判别分析的思想, 提出一种改进的 SVM 多分类方法 (D-SVM), 该方法不仅能够保留 SVM 较好的分类效果, 还能解决由于训练样本较少而对 SVM 产生的影响。同时, 本文还提出一种非参数变量筛选的方法, 选出对分类结果有显著作用的特征或变量, 从而达到降维和提高分类效果的作用。将本文提出的方法应用到 KDD99 数据集, 结果表明, 本文提出的方法具有高的分类精度。

2 方法介绍

2.1 SVM 方法介绍

SVM 的核心思想就是将样本映射到一个高

维空间, 并在高维空间中线性可分。再通过在高维空间构造一个超平面来达到分类的目的。SVM 一般用在解决二分类问题上, 具有良好的效果; 对于多分类问题, SVM 分类效果较差。为了使 SVM 在处理多分类问题时也能达到较好的分类效果, 本文采用二叉树分类思想, 首先将多分类问题处理成二分类问题, 再进行逐步迭代, 直至将所有的类都分开。

2.2 距离判别法

判别分析可以不受样本不均衡问题的影响, 分类效果比较理想且稳健。距离判别的核心思想就是通过计算待判样本到已知类中心的距离, 再比较距离的大小, 样本到哪个类中心的距离最小, 则该样本就判定属于哪一类。距离判别法具有简单高效的特点, 它通过计算距离, 直观地将样本归类。在某些方法失效的情况下, 距离判别仍具有比较稳健的分类效果。距离判别的定义可以表示为

$$\begin{cases} x \in G_1, & d(x, G_1) < d(x, G_2) \\ x \in G_2, & d(x, G_2) < d(x, G_1) \\ \text{unknown}, & d(x, G_1) = d(x, G_2) \end{cases} \quad (1)$$

其中, G_1 , G_2 表示类 1 和类 2; $d(x, G_1)$ 和 $d(x, G_2)$ 表示 x 到类 1 和类 2 的距离。在计算距离时一般选用马氏距离。相比于欧氏距离, 马氏距离不仅能够消除变量间量纲的影响, 还能消除多维变量间多重相关性的影响。

3 数据处理

3.1 KDD99 数据集介绍

KDD99 数据集是 1998 年美国国防部高级规划署在 MIT 林肯实验室进行的一项入侵检测评估项目。通过模拟真实网络环境, 仿真各种用户类型、各种不同的网络流量和攻击手段。收集了 9 周时间的 TCPdump 数据。随后来自哥伦比亚大学的 Sal Stolfo 教授和来自北卡罗莱纳州立大学的 Wenke Lee 教授对以上数据进行特征分析和数据预处理, 形成了一个新的数据集, 使其只包含网络流量数据。该数据集被用于 1999 年举行的 KDD CUP 竞赛中, 成为著名的 KDD99 数据集。KDD99 数据集是公认的入侵检测的 Benchmark 数据集。

KDD 数据集分为全部数据集 (kddcup.data.gz, 18M) 和 10%数据集 (kddcup.data_10_percent.gz, 2.1M)。本文以 10%数据集为分析对象。

10%数据集包含 494 021 条记录, 每条记录由 41 个特征变量和类别标签表示, 其中, 34 个是连续型变量, 7 个是名义型变量; 包含 5 类数据: Normal (97 278 条)、DoS 攻击 (391 458 条)、Probe 攻击 (4 107 条)、U2R 攻击 (52) 和 R2L 攻击 (1 126 条)。本文主要使用 34 个连续型数值变量作为研究的特征变量。

3.2 非参数检验变量筛选

变量筛选的目的是选出对分类结果有显著影响的特征变量, 从而可以达到降维, 提高效率的作用。不同于已有的变量综合的方法, 如 PCA, 本文提出一种利用非参数检验筛选变量的方法。

Kolmogorov-Smirnov 两样本分布检验, 从样本经验分布出发, 利用大样本性质检验 2 个样本是否来自于同一个总体。假定样本 $x_1, x_2, \dots, x_m$ 来自 $F(x)$ 分布, 样本 $y_1, y_2, \dots, y_n$ 来自 $G(y)$ 分布, 则 Kolmogorov-Smirnov 检验如式(2)所示。

$$\begin{aligned} H_0: & \forall x, F(x) = G(x) \\ H_1: & \exists x, F(x) \neq G(x) \end{aligned} \quad (2)$$

令 $F_m(x)$ 和 $G_n(y)$ 分别为 2 个样本的经验分布, 检验统计量 $D_{m,n}$ 如式(3)所示。

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| \quad (3)$$

Kolmogorov-Smirnov 检验已经运用到许多方面^[15-17], 在本文中, 利用 Kolmogorov-Smirnov 检验方法检验正常样本与入侵或攻击样本在哪些特征变量上分布具有显著的差异, 选出检验显著的特征变量作为研究的变量。与 PCA 等变量综合的方法不同, 本文提出的非参数检验变量筛选的方法可以直接选出正常样本与入侵或攻击样本在统计上有显著差异的变量, 删除不显著的变量, 消除冗余信息和不显著变量信息对分类的干扰, 比变量综合的方法更简便、有效。

4 实验结果及分析

4.1 Kolmogorov-Smirnov 检验变量选择

为了研究的简便, 本文以 KDD99 数据集的

10%数据集为研究对象, 且只选取 34 个连续型变量和标签变量作为研究的特征变量。因此, 得到了一个 $494\ 021 \times 35$ 的数据表, 其中表的每一行表示一条记录, 前 34 列表示选取的 34 个特征变量 (数值型), 第 35 列是标签, 表示每一条记录所属的类别 (Normal, DoS, Probe, R2L, U2R)。将 10%数据集分为正常、入侵或攻击 2 类, 分别记为类 N 和类 A, 其中, 类 N 中样本数为 97 278, 占 19.6%; 类 A 中样本数为 396 743, 占 80.4%。运用 Kolmogorov-Smirnov 方法检验类 N 和类 A 在哪些变量上的分布不存在显著差异, 结果如表 1 所示。

表 1 KS 检验不显著特征变量

变量	KS 检验
wrong_fragment	0.432 3
urgent	1.000 0
hot	0.474 0
num_failed_logins	1.000 0
num_compromised	0.060 2
root_shell	1.000 0
su_attempted	1.000 0
num_root	0.010 0
num_file_creations	0.796 8
num_shells	1.000 0
num_access_files	0.080 0
error_rate	0.722 5
srv_error_rate	0.104 5
num_outbound_cmds	1.000 0

注: 显著性水平为 0.01。

从表 1 中可以看出这 14 个特征变量在类 N 和类 A 上表现并无显著差异。因此, 可以认为这些特征变量对分类结果没有显著影响。最终得到了一个 $494\ 021 \times 21$ 的数据表。

4.2 正常与异常行为分类

对于入侵检测, 首先关心的是将正常行为和异常行为分开, 这是评价研究方法的关键指标。首先利用 SVM 分类方法构建第一级分类器将正常行为样本与异常行为样本分开。

在数据集中按比例随机选取 5 000 个正常样本, 20 000 个异常样本作为训练集。按照同样的

方式选取 10 000 个正常样本和 40 000 个异常样本作为测试集。独立重复进行 5 次实验，结果如表 2 所示。

表 2 5 次独立实验平均结果

类型	预测值		召回率/%	准确率/%
	Attack	Normal		
Attack	39 830	170	99.58	99.92
Normal	33	9 967	99.67	98.32

注：平均检测正确率为 99.59%。

从表 2 中可以看出，SVM 可以有效地将正常和异常样本分隔开，分类效果比较理想。正常行为和异常行为的分类准确率分别达到 99.92% 和 98.32%，整体分类正确率达到 99.59%。说明第一级分类可以以较高准确率（99.92%）识别入侵或攻击行为，为入侵行为的进一步分类奠定了基础。

4.3 异常行为具体攻击类别分类

4.3.1 R2L、U2R 与 DoS、Probe 分类

使用 KDD99 数据集研究多分类文献中，关于 R2L 和 U2R 的分类效果都不理想，原因在于这 2 类攻击样本太少，导致训练不足。考虑到数据集中 R2L 和 U2R 样本数较少，将这 2 类攻击合并为一类攻击，记作 R&U；DoS 和 Probe 当作一类，记作 D or P。利用 SVM 分类，结果如表 3 所示。

表 3 R&U 和 D or P 的平均误报率（5 次独立实验）

类型	预测值		误报率/%
	R&U	D or P	
R&U	56	44	44
D or P	11	40 389	0.027

注：平均检测正确率为 99.86%。

从表 3 中可以看出，即使使用二分类 SVM 方法，R&U 的误报率仍然很高，达到 44%，SVM 方法对于 R&U 的分类效果很差。

为了解决 R&U 误报率高的问题，随机选取 5 000 个正常样本，求出样本中心。按比例分别随机选取 40 400 个 D or P 样本和 100 个 R&U 样本，分别计算这 2 类样本到正常样本中心的马氏距离。这 2 类的马氏距离分布如图 1 所示。

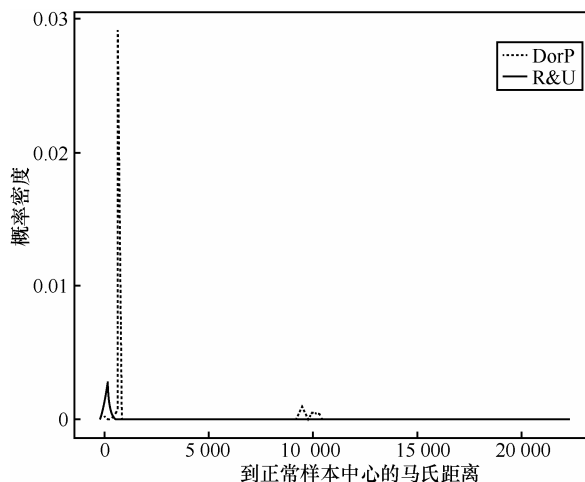


图 1 R&U 和 D or P 到正常样本类中心马氏距离分布

从图 1 中可以清楚地看到 R&U 和 D or P 这 2 类到正常样本中心的马氏距离具有明显的位置差异，说明这 2 类到正常样本中心的马氏距离显著不同。因此，可以用距离判别方法构建第二级分类器将这 2 类分开。

随机选取 5 000 正常样本，计算每一个样本到中心的马氏距离的均值和方差，分别用 MN 和 VN 表示。随机选取 40 400 个 D or P 样本和 100 个 R&U 样本，组成待分类样本记作 Test，分别计算 Test 中每一个样本到正常样本中心的马氏距离，用集合 $MdTest$ 表示。计算 $MdTest$ 中每个元素到正常样本中心的偏差程度，用 w 表示，如式(4)所示。

$$w = \frac{(MdTest - MN)^2}{VN} \quad (4)$$

为了确定 w 的最优取值，令 w 从 0 到 4 依次取值。每次增加 0.1，得到 R&U 的误报率如图 2 所示。

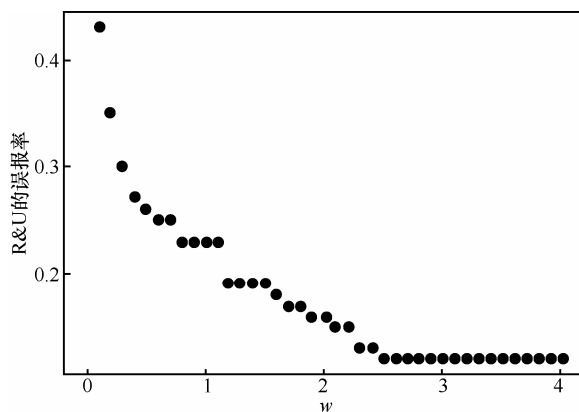


图 2 R&U 误报率与 w 变化关系

如图2所示, R&U的误报率随着 w 的增加, 呈现逐步下降的趋势, 两者之间表现出负相关关系。此外, 随着 w 的增加, D or P的误报率逐渐上升, R&U与D or P整体分类准确率呈下降趋势, 为了平衡R&U误报率和整体分类准确率, 本文选取图2中R&U的第3个平稳点作为 w 的最优取值($w=1.2$)。

为了测试距离判别方法的分类效果, 按照上文方法进行5次独立重复实验, 结果如表4所示。

表4 R&U与D or P平均误报率

类型	预测值		误报率/%
	R&U	D or P	
R&U	82	18	18
D or P	431	39 969	1.067

注: 平均检测正确率为98.89%。

对比表3和表4可以看出, 相比于SVM方法, 距离判别不仅能够大幅度降低R&U的误报率, 同时, 分类正确率较SVM降低得也比较小(R&U的误报率降低59.09%, 而检测正确率才只下降0.97%), 可以认为两者的分类正确性没有显著差别。由此可见, 距离判别方法较SVM方法在R&U与D or P分类问题上表现更好。

4.3.2 DoS与Probe分类

类R&U和类D or P充分分开后, 再使用SVM构建第三级分类器, 对DoS和Probe判别分类, 分类结果如表5所示。

表5 DoS和Probe分类结果(5次平均结果)

类型	预测值		准确率/%
	DoS	Probe	
DoS	39 996	4	99.95
Probe	20	380	98.96

注: 平均正确率为99.94%。

表5的分类结果表明, SVM方法可以有效地将DoS和Probe分开, 具有较高的分类准确率(DoS为99.95%, Probe为98.96%)和正确率(99.94%)。

以上结果表明, 采用三级分类器构建的SVM多分类方法, 即D-SVM方法, 既能保持SVM分

类的优势, 同时又能解决由于训练样本不足导致分类效果不理想的问题。

4.4 结果比较

为了比较D-SVM方法与其他方法的分类效果, 选取了PCA-Logit^[18]、CANN^[19], 结果如表6所示。

表6 D-SVM与PCA-Logit、CANN分类结果比较

方法	检测率/%				准确率/%
	Normal	DoS	Probe	U&R	
D-SVM	99.67	99.99	95.00	82	99.59
PCA-Logit	96.65	99.94	55.77	40.91	99.21
CANN	97.04	99.68	87.61	55.01	99.46

注: $DR = \frac{TP}{TP + FN}$, $Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$ 。

通过表6可以看出, 本文提出的D-SVM方法相比于其他2种方法具有一定的优势, 具有较高的整体分类准确率(accuracy)和检测率(DR)。在整体分类准确率指标上, D-SVM略高于其他2种方法; 但在检测率指标上, D-SVM具有显著的优势, 尤其体现在U2R和R2L的检测率上, D-SVM方法的U&R检测率达到82%, 远远高于其他2种方法, 从而说明, 即使在样本不均衡的情况下, D-SVM较其他2种方法仍然能够更加准确地识别出U2R和R2L 2类攻击。因此, 与其他2种方法相比, D-SVM具有较高的分类准确率和检测率; 同时还可以较少地受到来自不均衡问题的影响, 具有稳健性。

5 结束语

本文基于SVM方法, 提出一种综合判别分析的多分类方法——D-SVM。通过构建一个三级分类器, 实现了正常行为与入侵行为的准确分类以及入侵行为所属类别分类的功能。第一级分类器利用SVM方法将正常和异常行为分开; 第二级分类器利用距离判别分析方法将类R&U和类D or P分开; 第三级分类器再利用SVM方法将DoS和Probe分开。并用非参数检验的方法选取对分类结果具有显著影响的特征变量, 从而实现降维和提高检测效率的作用。KDD99数据集的实验结果表明, 本文提出的方

法具有较高的分类准确率和较低的误报率, 并克服了由于样本不均衡导致的训练不足、分类结果不理想的缺点。

参考文献:

- [1] LEE W, STOLFO S J, MOK K W. A data mining framework for building intrusion detection models[C]//The IEEE Symposium on Security and Privacy. c1999: 120-132.
- [2] VERWOERD T, HUNT R. Intrusion detection techniques and approaches[J]. Computer Communications, 2002, 25(15): 1356-1365.
- [3] ENDORF C F, SCHULTZ E, MELLANDER J. Intrusion detection & prevention[M]. McGraw-Hill Osborne Media, 2004.
- [4] LIAO H J, LIN C H R, LIN Y C, et al. Intrusion detection system: a comprehensive review[J]. Journal of Network and Computer Applications, 2013, 36(1): 16-24.
- [5] SHYU M L, CHEN S C, SARINNAPAKORN K, et al. A novel anomaly detection scheme based on principal component classifier[R]. Coral Gables Department of Electrical and Computer Engineering of Miami University, 2003.
- [6] JAMDAGNI A, TAN Z, HE X, et al. Repids: a multi tier real-time payload-based intrusion detection system[J]. Computer Networks, 2013, 57(3): 811-824.
- [7] 胡志鹏, 魏立线, 申军伟, 等. 基于核 Fisher 判别分析的无线传感器网络入侵检测算法[J]. 传感技术学报, 2012(2): 246-250.
HU Z P, WEI L X, SHEN J W, et al. An intrusion detection algorithm for wsn based on kernel Fisher discriminant[J]. Chinese Journal of Sensors and Actuators, 2012(2): 246-250.
- [8] MOK M S, SOHN S Y, JU Y H. Random effects logistic regression model for anomaly detection[J]. Expert Systems with Applications, 2010, 37(10): 7162-7166.
- [9] 郭春. 基于数据挖掘的网络入侵检测关键技术研究[D]. 北京: 北京邮电大学, 2014.
GUO C. Research on key technologies of network intrusion detection based on data mining[D]. Beijing: Beijing University of Posts and Telecommunications, 2014.
- [10] LEE W, STOLFO S J, MOK K W. Adaptive intrusion detection: a data mining approach[J]. Artificial Intelligence Review, 2000, 14(6): 533-567.
- [11] HWANG T S, LEE T J, LEE Y J. A three-tier IDS via data mining approach[C]//The 3rd Annual ACM Workshop on Mining Network Data. c2007: 1-6.
- [12] AN W, LIANG M. A new intrusion detection method based on SVM with minimum within class scatter[J]. Security and Communication Networks, 2013, 6(9): 1064-1074.
- [13] MUKKAMALA S, JANOSKI G, SUNG A. Intrusion detection using neural networks and support vector machines[C]//The 2002 International Joint Conference on Neural Networks(IJCNN'02) c2002: 1702-1707.
- [14] GAN X S, DUANMU J S, WANG J F, et al. Anomaly intrusion detection based on PLS feature extraction and core vector machine[J]. Knowledge-Based Systems, 2013, (40): 1-6.
- [15] 叶钢, 余丹, 李重文, 等. 一种基于 Kolmogorov-Smirnov 检验的缺陷定位方法[J]. 计算机研究与发展, 2013(4): 686-699.
YE G, YU D, LI C W, et al. Fault localization based on Kolmogorov-Smirnov testing model[J]. Journal of Computer Research and Development, 2013(4): 686-699.
- [16] 从飞云, 陈进, 董广明. 基于 AR 模型的 Kolmogorov-Smirnov 检验性能退化及预测研究[J]. 振动与冲击, 2012(10): 79-82.
CONG F Y, CHEN J, DONG G M. Performance degradation assessment by Kolmogorov-Smirnov test and prognosis based on AR model[J]. Journal of Vibration and Shock, 2012(10): 79-82.
- [17] 陈敏. 门限自回归模型条件异方差的 Kolmogorov-Smirnov 检验[J]. 应用数学学报, 2002(4): 577-590.
CHEN M. A Kolmogorov-Smirnov test of conditional heteroscedasticity for threshold autoregressive models[J]. Acta Mathematicae Applicatae Sinica, 2002(4): 577-590.
- [18] 李蕊. 基于 PCA 和 LOGIT 模型的网络入侵检测方法[J]. 成都信息工程学院学报, 2014(3): 261-267.
LI R. A network intrusion detection method based on PCA and LOGIT model[J]. Journal of Chengdu University of Information Technology, 2014(3): 261-267.
- [19] LIN W C, KE S W, TSAI C F. CANN: an intrusion detection system based on combining cluster centers and nearest neighbors[J]. Knowledge-based systems, 2015, 7(8): 13-21.

作者简介:



张辉(1979-), 女, 河南镇平人, 硕士, 新疆公安厅特别侦察队技术八级工程师, 主要研究方向为网络与信息安全、网络侦察技术。



刘成(1985-), 男, 湖南邵阳人, 博士, 国家计算机网络应急技术处理协调中心高级工程师, 主要研究方向为网络与信息安全、网络攻防技术。