# 机器学习在网络空间安全研究中的应用

张蕾[1) ] 崔勇[1)] 刘静[2)] 江勇[1)] 吴建平[1)]

[1)](清华大学 计算机科学与技术系 北京 100084)

[2)](北京邮电大学 网络与交换技术国家重点实验室 北京 100876)

**摘 要** 随着云计算、物联网、大数据等新兴技术的迅猛发展，数以亿计的网络接入点、联网设备以及网络应用产生的海量数据，给网络空间安全带来了巨大的困难和挑战，传统的安全问题解决方案面对海量数据变得效率低下。机器学习以其强大的自适应性、自学习能力为安全领域提供了一系列有效的分析决策工具，近年来引起了学术界与工业界的广泛关注和深入研究。为此，本文以网络空间安全一级学科为指导，围绕机器学习技术应用于网络空间安全领域的最新研究成果，首先详细阐述了机器学习技术在网络空间安全研究中的应用流程；然后从系统安全、网络安全和应用安全三个层面，着重介绍了机器学习在芯片及系统硬件安全、系统软件安全、网络基础设施安全、网络安全检测、应用软件安全、社会网络安全等网络空间安全领域中的解决方案，重点分析、归纳了这些解决方案中的安全特征及常用机器学习算法。最后总结了现有解决方案中存在的问题，以及机器学习技术在网络空间安全研究中未来的发展方向和面临的挑战。

关键词 网络空间安全；机器学习；系统安全；网络安全；应用安全

**中图法分类号** TP393

## Application of Machine Learning in Cyberspace Security Research

ZHANG Lei[1)] CUI Yong[1)] LIU Jing[2)] JIANG Yong[1)] WU Jian-Ping[1)]

[1)](Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

[2)](State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876)

**Abstract** With the rapid development of new technologies such as cloud computing, Internet of Things and big data, hundreds of millions of network access points, networking devices and network application, as well as the massive data which they generate, bring great difficulties and challenges to cyberspace security. Under the circumstance, some traditional solutions to security problem have become inefficient, while machine learning can be efficient by providing a series of effective analysis and decision-making tools because of its strong adaptability, self-learning ability. In recent years, both academia and industry have been attracted by cyberspace security based on machine learning, and have gained a certain research achievements. Therefore, we use the first level discipline of cyberspace security as a guide and review the latest research achievements about cyberspace security based on machine learning, mainly including chip security, hardware security, system software security, network infrastructure security, network security defense and protection, application software security, social

本课题得到国家自然科学基金项目(No.61422206))资助.张 蕾，女，1979年生，博士研究生，主要研究领域为机器学习与网络安全.E-mail: zhanglei_ujn@126.com.崔 勇(**通信作者**)，男，1976年生，博士，教授，博士生导师，中国计算机学会(CCF)高级会员,主要研究领域为计算机网络体系结构、移动互联网、网络空间安全.E-mail: cuiyong@tsinghua.edu.cn. 刘 静，女，1993年生，硕士研究生，主要研究领域为机器学习、网络安全与隐私保护.江 勇，男，1975年生，博士，教授，博士生导师，中国计算机学会(CCF)高级会员，主要研究领域为计算机网络体系结构、下一代互联网、移动互联网.吴建平，男，1953年生，中国工程院院士，教授，博士生导师，主要研究领域为计算机网络体系结构、下一代互联网、网络空间安全.

network security. And we elaborate its workflow, especially security features and common machine learning algorithms. At last, the issues, some future research trends and challenges are explored.

**Key words** cyberspace security; machine learning; system security; network security; application security

## 1 引言

网络空间（Cyberspace）不仅包含互联网、通信网、各种计算系统、各类嵌入式处理器和控制器等硬件和软件，也包括这些硬件和软件产生、处理、传输、存储的各种数据或信息，还包括人类在其中活动而产生的影响。网络空间因而被称为陆、海、空、太空之外的第五大空间[1]。近年来网络空间中各类安全事件和网络攻击频繁发生，例如 2016 年 10 月由恶意软件 Mirai 控制的僵尸网络发起 DDoS 攻击，造成美国东海岸大范围断网；2017 年 5 月爆发的勒索病毒软件 WannaCry 利用系统漏洞进行攻击，造成全球多个国家数十万用户电脑中毒；在我国，每年因伪基站、恶意软件勒索等数字犯罪造成的损失达上百亿元。上述事例表明，网络空间的安全不仅影响着国民经济的发展，还关系着社会的稳定和国家安全，因此网络空间安全（Cyberspace Security，又称 Cyber Security）受到政府、学术界及工业界的高度重视。

我国于 2015 年正式批准设立“网络空间安全”国家一级学科，目前网络空间安全的研究主要涉及五个研究方向，即网络空间安全基础、密码学及应用、系统安全、网络安全、应用安全[2]。其中，系统安全主要研究网络空间中的单元计算系统的安全；网络安全主要研究网络自身和传输信息的安全；应用安全则研究各类应用系统和综合应用的安全；密码学及应用为系统、网络及应用安全提供密码安全机制；网络空间安全基础则为其他方向提供理论、架构和方法学。

随着承载人们工作、生活的移动互联网、云计算、大数据、物联网、机器学习等技术的迅猛发展，网络空间安全所面临的环境日益复杂，面对的安全威胁也日益升级。在这种复杂环境下，传统的以分析安全问题、固定规则设定的研究方法变得效率低下、甚至无能为力。例如依靠安全专家人工修复方法无法解决零日漏洞问题；传统依靠固定规则的网络入侵检测方法，面对不断增大的数据维度和复杂的网络行为，出现大量误判警告或判别时间较长；依靠固定规则或黑白名单过滤的垃圾邮件检测方法存在检测效率低，规则更新不及时等问题。随着网络空间安全问题的复杂度越来越高，数据维度不断增加，对网络空间安全问题的研究提出了新的需求。近年来，机器学习在计算机视觉、语音识别、自然语言处理、医疗数据分析等方面的应用取得了瞩目的研究成果，展现了机器学习在解决分类、预测以及辅助决策中强大的能力。机器学习技术为解决传统方法难以建模的网络空间安全问题提供了可能性。

在 20 世纪 80 年代，已有学者在网络入侵检测中应用机器学习技术[3]，但受限于当时的存储空间及计算能力，机器学习未能引起学者们的重视。随着大数据、云计算技术的出现，对搜集、存储、管理及处理数据的能力大幅度提升，因此，将机器学习应用到网络空间安全中，已成为近年来安全领域的研究热点。安全领域四大顶级会议（CCS、S&P、USENIX、NDSS）近年来收录了 50 余篇机器学习在网络空间安全中的相关研究工作，CCS 会议甚至成立了专题 AISec 研讨人工智能技术在安全和隐私方面的应用；此外，在汇聚世界各地信息安全从业人员的美国黑帽大会上，机器学习和安全领域结合的议题也成为关注热点以及未来趋势之一。

目前，已有学者对机器学习与网络空间安全中的部分安全问题进行了研究梳理和总结。Jiang 等人[4]从方法、算法和系统设计三个方面系统总结了 2008 年-2016 年之间应用机器学习到部分安全问题的研究工作，主要包含网络安全、安全服务、软件和应用安全、系统安全、恶意软件、社会工程以及入侵检测系统这 7 个安全研究方向。Buczak[5]、Sommer[6]等人从模型的构建及部署问题介绍了入侵检测和机器学习结合的研究工作，Nishani[7]等人则是介绍了针对移动自组织网络入侵检测和机器学习结合的研究。总体而言，上述综述论文在内容上或仅针对机器学习在某一子领域应用进行研究，或侧重于理论分析研究。

本文针对网络空间安全一级学科的五个研究

方向进行调研, 调研结果显示，除了边信道攻击研究外，机器学习在网络空间安全基础、密码学及其应用作为理论基础方面的研究较少涉及；而在系统安全、网络安全、应用安全三个方向中有大量的研究成果发表。其中，系统安全以芯片、系统硬件物理环境及系统软件为研究对象，网络安全主要以网络基础设施、网络安全检测为研究重点，应用层面则关注应用软件安全、社会网络安全。

本文对近年来机器学习在网络空间安全研究中的应用成果进行归类和梳理，形成如图 1 所示的研究体系。从机器学习技术应用于网络空间安全的角度出发，本文总结了机器学习一般应用流程，如图 1 中右侧所示，详细介绍问题的定义、数据采集、数据预处理及安全特征提取以及模型构建、验证、效果评估各个阶段，有助于研究人员全面的理解基于机器学习技术的网络空间安全问题解决方案。接下来本文从系统安全、网络安全以及应用安全三个研究领域，按照上述机器学习的一般流程，对机器学习在网络空间安全领域的典型应用进行了分析和讨论，典型应用如图 1 中左侧所示，着重综述了现有研究成果的技术思路，同时针对安全问题抽象、安全数据的使用、常用安全特征及机器学习算法的选择等机器学习关键技术进行详细阐述。本文最后还对存在问题、未来发展趋势及挑战进行了总结与分析。

本文组织结构如下：第 2 节剖析了机器学习在网络空间安全中的应用流程。第 3、4、5 节分别对机器学习在系统安全、网络安全和应用安全中的研究进行了梳理和总结。第 6 节从不同层面分别对机器学习在网络空间安全中的应用进行了展望，并分析了所面临的挑战。第 7 节概括总结了全文工作。

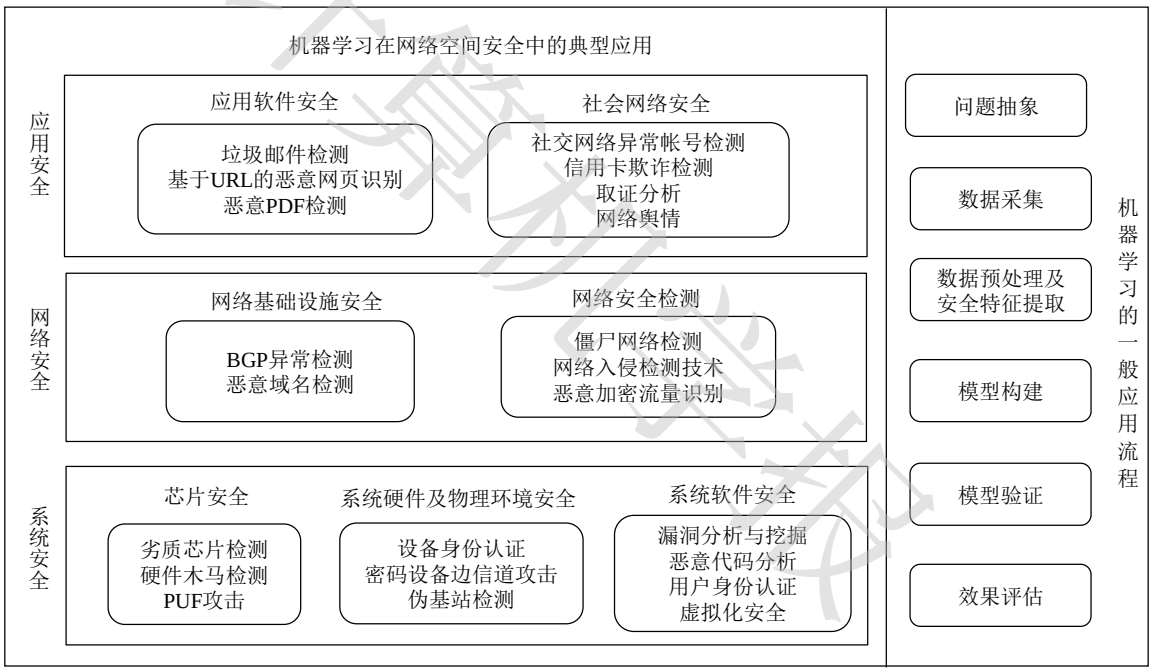


图 1 机器学习在网络空间安全研究中的应用及流程

## 2 机器学习在网络空间安全中的应用流程

通常机器学习被认为是一组能够利用经验数据来改善系统自身性能的算法集合[8]。机器学习从大量数据中获取已知属性，解决分类、聚类、降维等问题。理解机器学习在网络空间安全中的应用流程，能够有效的帮助网络空间安全领域的研究人员建立直观的认识，同时也是其进一步采用机器学习技术解决网络空间安全问题的前提。如图 2 所示，机器学习在网络空间安全研究中的一般应用流程，主要包括安全问题抽象、数据采集、数据预处理及安全特征提取、模型构建、模型验证以及模型效果评估 6 个阶段。在整个应用流程中，各阶段不能独立存在，相互之间存在一定的关联关系。本节按照机器学习在网络空间安全中的一般应用流程，详细阐述每个阶段代表的含义及典型的安全实例。

### 2.1 安全问题抽象

安全问题抽象是将网络空间安全问题映射为机器学习能够解决的类别。问题映射恰当与否直接关系着机器学习技术解决网络空间安全问题成功与否。因此，使用机器学习技术解决安全问题的第一步就是要进行问题的抽象和定义，将安全问题映射为机器学习能够解决的分类、聚类及降维等问题。如图 3 所示，对劣质芯片或硬件木马的检测、伪基站检测、虚拟化安全、信用卡欺诈等都可以抽象为分类问题；设备身份认证、社交网络异常帐号检测、网络入侵检测等可以抽象为聚类问题；用户身份认证、 恶意/异常/入侵检测、取证分析、网络舆情等既可以抽象为分类问题也可以抽象为聚类问题。如果是高维数据的处理，可以抽象为降维问题，例如在设备身份认证、恶意网页识别问题中，由于数据维度过高，可以利用机器学习主成分分析（Principal Component Analysis，PCA）算法、奇异值分解（Singular Value Decomposition，SVD）算法等对数据进行降维操作。通过对安全问题的合理抽象和定义，研究人员可以明确如何采集数据，并选择恰当的机器学习算法构建安全问题模型。

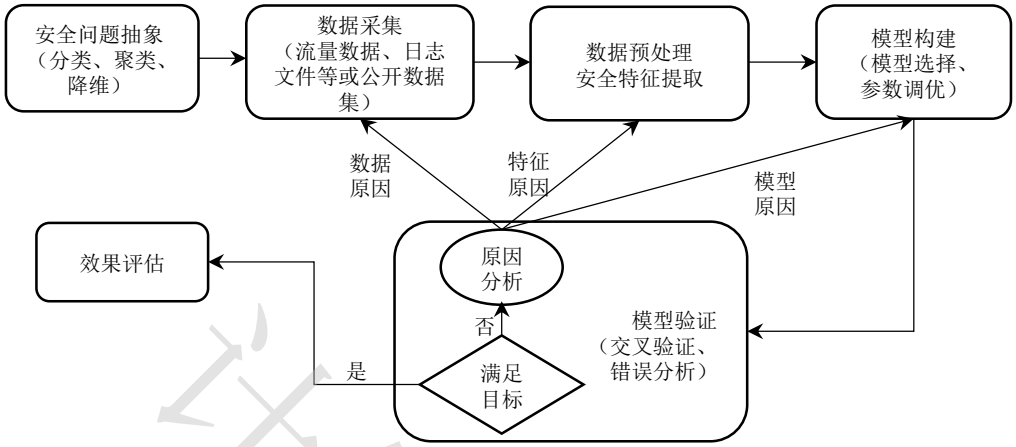


图 2 机器学习在网络空间安全研究中的应用流程

### 2.2 数据采集

应用机器学习算法必不可少的要有大量的有效数据，因此数据采集是机器学习应用于网络空间安全的前提条件。数据采集阶段主要利用各种手段，如 Wireshark、Netflow、日志收集工具等，从系统层、网络层及应用层采集数据。系统层数据用于系统安全问题的研究，这类数据主要有芯片信息、设备信息、系统日志信息以及实时运行的状态信息等，主要用于芯片安全、设备安全及系统软件安全，例如采集基站的位置信息、短信日志等数据用于伪基站检测研究[9]。网络层数据指与具体网络活动密切相关的数据，目前常用的是网络包数据或网络流数据，主要用于检测僵尸网络、网络入侵等，例如在企业内部网络中采集大量的真实的 TCP 流数据用于进行协议分类及异常协议检测研究[10]。应用层数据指网络空间中的各类应用软件产生及存储的数据，如邮件文本信息、web 日志、社交网络文本信息、用户个人信息等，主要用于应用软件安全检测、网络舆情分析等，例如采集大量的 URL 数据用于恶意网页识别[11]。

除自行采集数据外，目前安全领域有一些常用的公开数据集供研究者使用，如表 1 所示。

表 1 公开数据集

| 序号 | 数据集名称 | 说明 |
|---|---|---|
| 1 | DARPA Intrusion Detection Data Sets[12-14] | 网络入侵检测数据集（包含 1998、1999、2000 三个数据集） |
| 2 | KDD[15] | 网络入侵检测数据集（包含 1998、1999 两个数据集） |
| 3 | UCI’s Spambase [16] | 垃圾电子邮件数据集 |
| 4 | Honeynet Project Challenges [17] | 网络攻击行为数据集 |
| 5 | Internet Traffic Archive[18] | 网络包数据集，包含路由信息 |
| 6 | Alexa 网站域名[19] | Alexa.com 收集的知名网站域名 |
| 7 | DMOZ Open Directory Project[20] | URL 地址集 |
| 8 | RIPE RIS 和 Route Views [21] | 域间路由数据集 |
| 9 | PhishTank[22] | 钓鱼网站 URL 地址集 |

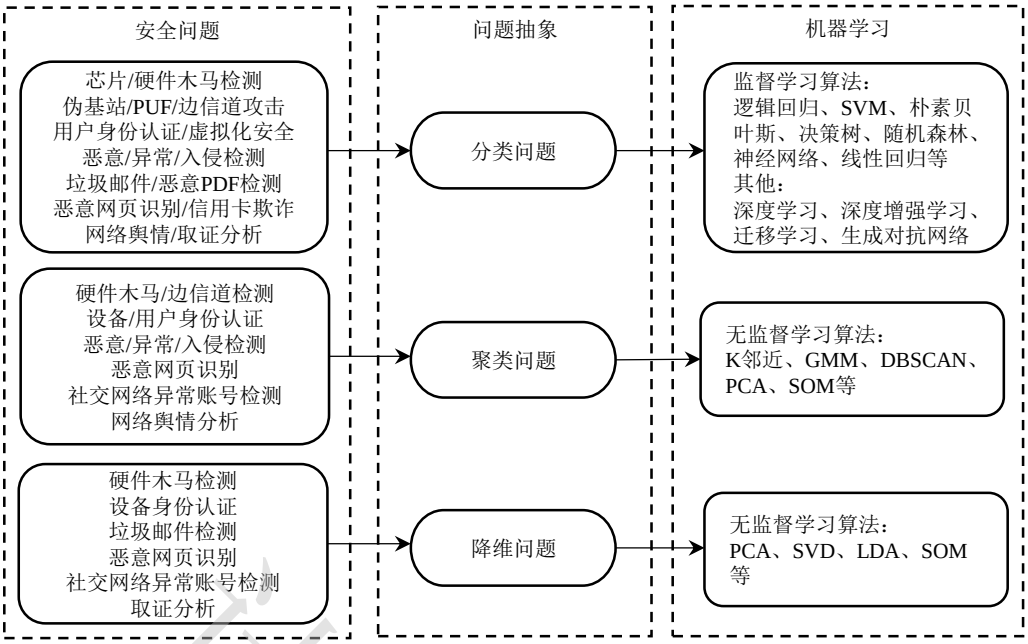


图 3 安全问题抽象

### 2.3 数据预处理及特征提取

由于采集的原始数据存在数据缺失、非平衡、格式不规范、异常点等问题，需要在提取特征之前对原始数据进行清洗和处理，主要包含对数据规范化、离散化以及非平衡性的处理等。

（1）数据预处理

在真实的网络环境中，采集的数据可能包含大量的缺失值、噪音，也可能由于人工录入失误而产生异常点。因此，为了提高数据的质量，保障构建模型学习的效果，需要对数据进行清洗及归一化等预处理[8]。数据预处理一般流程是先分析统计数据，然后对缺失值、异常值、重复值、噪音数据等进行清洗，清洗之后对数据进行归一化操作。例如从企业内部采集的 TCP 流数据，首先需要剔除重复数据[10]、去除噪音等规范化操作；然后对清洗之后的数据进行聚合[12]、归一化等处理[11][23][24]。

（2）数据缺失处理及异常值的处理

如果采集数据集中某个特征缺失值较多时，通常会将该特征舍弃，否则可能会产生较大的噪声，影响机器学习模型的效果。当某个特征的缺失值较少时，可用采用固定值填充、均值填充、中位数填充、上下数据填充、插值法填充或者随机数填充等方法。此外，有些机器学习模型诸如随机森林、贝叶斯网络及人工神经网络等，自身能够处理数据缺失的情况，不需要对缺失数据做任何的处理。

如果采集的数据集中存在的异常值，通常采用直接将该条数据删除，或将其视为缺失值处理。

（3）非平衡数据的处理

在信用卡欺诈检测[25]、网络入侵检测[6]、移动终端恶意代码检测[26]等安全问题中使用的数据集，异常数据样本或恶意数据样本远远少于正常样本。对于这种非平衡的数据集，直接使用机器学习算法构建检测模型效果往往不佳。为了解决非平衡数据问题，通常使用过采样或欠采样方法[8]构造平衡数据集。欠采样是当数据量足够大时，通过减少多数类，即数据量占大多数的类别的样本量来平衡数据集，例如在信用卡欺诈检测中，减少正常账号的数据样本。过采样与欠采样相对，过采样适用于数据量不足的情况，通过复制、自举法等方法增加少数类的样本量来平衡数据集，例如增加伪造信用卡账号的数据样本。

（4）数据集的分割

数据预处理完成后，还要进行机器学习模型所需数据集的准备工作。该工作主要将整理之后的数据集分为三个集合：训练集、验证集和测试集。训练集用于机器学习模型的构建，验证集用于验证模型及参数调优，测试集用于评估模型在实际使用中的泛化能力。常用的数据集分割方法有随机抽样、交叉验证法等[8]，例如网络入侵检测研究中，常采用随机抽样的方法分割数据集，即随机选择 90%的数据作为训练集，随机选择 10%的数据作为测试集，然后采用交叉验证法分割为训练集和验证集[27]。

（5）特征提取

特征提取指从数据中提取最具有安全问题的本质特性的属性。从清理后的数据中提取特征通常需要特定的领域知识，例如恶意网页的识别中需要从抓取的网页数据中提取主机信息特征、网页内容特征、静态链接关系及动态网页行为等特征[28]。这些特征提取通常由具有一定领域知识的专业人员完成，这种人工完成特征提取的方式比较困难而且费时。近年来机器学习中新兴的深度学习是进行自动提取特征的一个很好的选择[29]。

### 2.4 模型构建

模型构建是机器学习在网络空间安全应用流程中的中心环节，根据数据预处理后的数据集及目标问题类型，在本阶段选择合适的学习算法，构建求解问题模型。模型构建具体包含 2 个部分，即算法选择和参数调优。需要面对种类繁多的机器学习算法，如何能够选择恰当的机器学习算法是应用机器学习技术解决网络空间安全问题的关键。

在机器学习领域，按照数据集是否有标记分为监督学习、无监督学习。在监督学习模式中，每组数据有一个明确的标签，例如垃圾邮件检测中的每条数据标记为“垃圾邮件”或“非垃圾邮件”。监督学习算法常用于分类问题和回归问题。常见算法有逻辑回归（Logistic Regression，LR）、人工神经网络（Artificial Neural Network，ANN）、支持向量机（Support Vector Machine，SVM）、决策树、随机森林、线性回归等。在非监督学习中，数据不包含标签信息，但可以通过非监督学习算法推断出数据的内在关联，例如社交网络帐号的检测中对好友关系、点赞行为等聚类[30]，从而发现帐号内在的关联。非监督学习常用于聚类问题。常见的算法有 K-Means 算法、k 近邻(k-Nearest Neighbor，KNN）、基于密度的 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)算法、层次聚类（Hierarchical Clustering）算法、图聚类算法等。

此外，近年来出现的深度学习、迁移学习、深度增强学习算法以及生成对抗网络也为解决网络空间安全问题提供了新的选择。深度学习[29]凭借强大的自动提取特征的能力，被用于解决异常协议检测[10]、恶意软件检测[31][32]、网络入侵检测[33]、BGP 异常路由检测[34]以及差分隐私保护[35-36]等安全问题。擅长于场景或领域迁移的迁移学习[37]在网络空间安全也有其用武之处，例硬件木马检测中利用迁移学习对边信道信号检测进行校正[38]。深度学习与增强学习相结合的深度增强学习算法（DQN）[39]应用于移动终端恶意软件检测[40]。生成对抗网络作为一种生成式模型[41]，和深度学习算法相结合用于随机域名生成算法[19]以及恶意代码检测[42]。

将选定的算法和训练数据集用于模型训练时，往往需要面临参数调优的巨大挑战。参数调优往往与训练目标、选择的算法相关，但目前参数调优的过程缺乏足够的理论指导，需要在庞大的参数空间来寻找可接受的参数或者依据个人经验进行调整。

### 2.5 模型验证

模型验证主要评估训练的模型是否足够有效。在此阶段中，k 倍交叉验证法是最常见的验证模型稳定性的方法[15,23,43-44]。k 倍交叉验证法将数据预处理后的训练数据集划分成 k 个大小相似且互斥的子集，每个子集尽可能保持数据分布的一致性，然后每次用 k-1 子集的并集作为训练集，剩余子集作为验证集，从而获得了 k 组训练数据集和验证集，可进行 k 次训练和验证测试，最终的返回结果是这 k 次验证测试结果的均值。例如在设备身份认证[45]、网络入侵检测[15]、恶意域名检测系统[46]、恶意 PDF 文件的检测[47]、社交网络异常帐号检测[44]中均使用了 10 倍交叉验证模型，用于评估模型是否符合训练目标。

如果当前模型与训练目标偏离较大，则通过分析误差样本发现错误发生的原因，包括模型和特征是否正确、数据是否具有足够的代表性等。如果数据不足，则重新进行数据采集。如果特征不明显，则重新进行特征提取；如果模型不佳，则选择其他学习算法或进一步调整参数。

### 2.6 效果评估

机器学习的模型评估主要关注模型的学习效果以及泛化能力。泛化能力的评估通常是对测试集进行效果评估。同时，由于不同领域有不同指标的提法，因此，本文仅针调研论文中所涉及到的有关分类问题和聚类问题的效果评估指标加以说明。

在芯片检测[48]、恶意软件检测[49]、异常检测[50]、网络入侵检测[51]等分类问题中，效果评估常用到表 2 所列的评估指标[52]，常用的分类评估指标有正确率、查准率（又称精度）和查全率（又称召回率）。正确率是分类正确的正常样本与恶意样本数占样本总数的比例，一般来说正确率越高，分类器越好。查准率则是被正确识别的正常样本数占被识别为正常样本的比例，也是分类器精确性的衡量标准。

查全率是被正确识别的正常样本与正确识别的正

表 2 安全领域分类问题效果评估指标

| 评估指标 | 描述 |
|---|---|
| TP (True Positive) | 被正确的识别为正常的样本数 |
| TN (True Negative) | 被正确的识别为恶意的样本数 |
| FP (False Positive) | 被错误的识别为正常的样本数 |
| FN (False Negative) | 被错误的识别为恶意的样本数 |
| ACY (Accuracy) | 正确率，*ACY=（TP+TN）/(TP+TN+FP+FN)* |
| P (Precision) | 查准率或精度，*P=TP/(TP+FP)* |
| TPR (TP Rate) | 查全率或召回率，*TPR=TP/(TP+FN)* |
| FPR (FP Rate) | 误报率，*FPR=FP/(FP+TN)* |
| FNR (FN Rate) | 漏报率，*FNR=FN/(TP+FN)* |

常样本和错误识别的恶意样本之和的比例，该指标衡量了分类器对正常样本的识别能力。此外，在不同的领域还有不同的指标说法，例如在硬件木马检测、异常检测、网络入侵检测中还常使用误报率（False Positive Rate，FPR）、漏报率（False Negative Rate，FNR）来衡量模型的泛化能力。在认证领域常使用误识率（False Acceptance Rate，FAR）、拒识率（False Rejection Rate，FRR）对模型进行效果评估。

在社交网络帐号检测[30]、异常检测[53]等聚类问题中，模型的目标是同一簇的样本尽可能彼此相似，不同簇的样本尽可能不同，因此聚类问题常用的评估指标分为两类[8]，一类是将聚类结果与某个参考模型进行比较；另一类是直接考察聚类结果而不利用任何参考模型。

# 3 机器学习在系统安全研究中的应用

网络空间中的系统主要指具有独立计算能力的单元计算系统，例如计算机、移动终端等。本节以这些单元计算系统的安全为核心，横跨芯片、系统硬件及物理环境、系统软件三个层面，介绍机器学习在系统安全中的相关研究。其中，芯片安全方面包括劣质芯片检测、硬件木马检测及 PUF 攻击；系统硬件及物理环境安全包括设备身份认证、物理层边信道攻击及伪基站检测；系统软件安全包括漏洞分析与挖掘、恶意代码分析、用户身份认证及虚拟化安全。

## 3.1 芯片安全

分布多维化、步骤繁多的硬件产业供应链[54]使硬件设备易在各个环节容中出现安全问题，例如二手芯片、硬件木马。已有学者尝试利用机器学习技术来解决芯片安全问题，主要基于边信号分析、指纹识别和图像识别的劣质芯片和硬件木马检测。针对芯片知识产权保护安全，已有研究提出物理不可克隆函数（PUF）攻击，主要是利用机器学习技术推测由 PUF 生成的芯片知识产权保护标识。

### 3.1.1 劣质芯片检测

劣质芯片包括低规格、不达标的芯片以及翻新的芯片。劣质芯片一般很难通过肉眼看出，传统检测方法例如物理检测中的材料分析、电子检测中的功能测试及结构测试等，代价昂贵又十分耗时[55]。研究发现，劣质芯片与原厂芯片存在差异参数：一是边信道差异参数，包括正偏压温度不稳定性、负偏压温度不稳定性、热载流子注入、路径延迟等；二是芯片外形方面，例如颜色、擦痕等。因此，出现了基于上述两类差异参数的劣质芯片检测研究。

基于边信道差异参数检测研究中，Huang 等人[56]选取了原厂芯片的若干差异参数作为正样例数据，使用单类支持向量机（One Class SVM，OC-SVM）训练劣质芯片分类器，从而实现在待测芯片中检测劣质芯片，有效减少了检测芯片的成本和时间开销。Xiao 等人[57]利用时钟扫描技术生成芯片路径延迟指纹，分别使用简单异常点分析（Simple Outlier Analysis，SOA）和主成分分析法（Principal Component Analysis，PCA）检测劣质芯片，PCA 算法识别准确率优于 SOA。该方案不需要使用额外的辅助电路，但不适合老化时间短且老化影响小的情况。

基于芯片外形的检测研究中，研究人员提出利用芯片图像特征识别劣质芯片。例如 Asadizanjani 等[58]以众包方式收集劣质芯片数据，基于其构建的开放劣质芯片图片数据库，再利用人工神经网络（Artificial Neural Network，ANN）算法提取待测芯片图片的特征，自动分类芯片。该方法的缺点是当芯片外表颜色、擦痕等变化不明显时，无法检测出劣质芯片。

劣质芯片检测目前使用的机器学习技术主要是单类分类器（如 OC-SVM）、异常检测技术（如 SOA）等，主要因为多数训练样本只有一类可信芯片样本[56]。劣质芯片检测不管是依据边信道差异参数特征还是利用图像识别技术，在一定程度上均提

升了检测效率，但这仅是在粗粒度差异特征下取得的效果，面对细粒度的差异特征时，劣质芯片识别率较低。

3.1.2 硬件木马检测

芯片的硬件木马通常指在原始芯片植入具有恶意功能的冗余电路[59]。硬件木马通常分为物理上的木马（例如增加或删除晶体管、开关选择器、连接线等）和激活态的木马（例如触发器和负载）。被植入木马的芯片，其热量、功耗和延时等边信道信号会有所改变，因此可以通过收集芯片边信道的参数指纹，在多维的空间对比，判断是否在可信芯片的参数指纹范围内[60]。目前机器学习在硬件木马检测方向研究有芯片原理图成像识别和边信道信号分析。

芯片原理图成像识别研究中，Bao 等人[61]提出在逆向工程单层成像后，将实验芯片与可信芯片的成像图差值作为特征参数，使用 OC-SVM 算法进行检测。在公开的基准测试中，该方法检测芯片硬件木马有着较高的准确率，但当芯片成像图网格参数较小以及木马由较小参数篡改构造时，较难检测出硬件木马。

芯片原理图成像技术需要昂贵的设备支持，并且减少芯片层级会对芯片造成毁灭性破坏。因此，较多的研究在分析边信道信号方向上。例如 Liu[38] 利用可信芯片电路仿真、蒙特卡洛分析得到多维边信道信号数据，再利用 PCA 算法对数据作降维处理，然后通过非线性回归模型得到边信道指纹，最后通过 OC-SVM 算法进行分类识别。此外，Liu 还利用 KMM 核均值匹配的参数迁移学习算法对边信道信号进行校正，以及利用基于统计分析的参数建模方法对蒙特卡洛仿真电路进行校正。Iwase[62]将可信芯片和植入硬件木马的芯片的能耗差值的时域图，通过离散傅里叶变换变换到频域，然后利用 SVM 算法学习能耗差值的频域特征，从而识别出带有硬件木马的芯片。Jap 等人[48]则在真实的 FPGA 密码设备上，利用 OC-SVM 算法构建了硬件木马检测平台，并做了 8 位和 128 位最低有效位的两组实验，与基于模板的硬件木马检测方法对比，发现利用 OC-SVM 检测硬件木马的准确率和误报率均更优。

总体而言，芯片原理图成像识别、边信道信号分析虽然都能在一定程度上检测出硬件木马。然而，它们的前提是硬件木马对芯片的电路或边信道参数有明显的改变，非常小的改变或深度隐藏的硬件木马将很难被检测出。

3.1.3 PUF 攻击

物理不可克隆函数（Physical Unclonable Function，PUF）[63]电路是一种根据芯片在制造过程中的差异性产生独特的激励－响应对（Challenge-Response Pairs）的电路。通过这种 PUF 电路可生成该芯片唯一的标识，通常这种标识很难被复制，因此 PUF 电路生成的光学 PUF、涂层 PUF、硅 PUF 等标识被用于保护芯片知识产权。然而，已有研究利用机器学习对 PUF 生成的标识进行攻击。

2010 年 Rührmair 等人[64]利用逻辑回归、SVM 和进化策略攻击了 Arbiter 仲裁器 PUF、环形振荡器（Ring Oscillator）PUF、异或门仲裁器 PUF、轻量级安全 PUF 和前向反馈 PUF。其攻击原理是收集给定 PUF 的激励响应对，利用机器学习算法推测任意激励下该 PUF 的响应值。针对各类 PUF，作者在不同电路位数、机器学习算法下做了攻击实验，均取得了较高的准确率，并得出结论：所有的强类型 PUF 均不安全，其他类型 PUF 可通过增加电路位数、电路设计复杂性来增强安全性。此外，Hospodar[65]仅针对 64 位 Arbiter 仲裁器 PUF 产生的激励响应对，使用人工神经网络和 SVM 进行硅 PUF 攻击模型的有效性测量。实验显示，当训练集为 500 个激励-响应对时，产生了 90%正确率的 PUF 攻击模型，当训练集为 5000 个激励-响应对时，产生了更高准确率的 PUF 攻击模型。这表明训练数据越大，PUF 攻击成功率越高。

上述 PUF 攻击研究表明基于 PUF 的芯片知识产权保护存在极大的漏洞，需要进一步增强 PUF 的安全性，例如增加 PUF 电路位数、电路设计复杂性[70]等。

3.2 系统硬件及物理环境安全

硬件设备身份认证是系统硬件常见的安全问题，例如网络设备传统的身份认证方式是依据 MAC 地址进行认证，这种方式很容易被伪造，因此出现基于硬件自身指纹特征的设备身份认证技术。本节主要介绍机器学习在设备身份认证技术中的应用，主要有暂态信号、调制信号和频谱响应这三类指纹特征。此外，在网络空间物理环境中，系统硬件与外部设备进行信息交换或通信时常常会遇到信息泄露、中间人攻击，例如常见的物理层边信道攻击、伪基站。因此本节还介绍了机器学习在物理层边信道攻击和伪基站检测中的研究。

3.2.1 设备身份认证

基于机器学习的设备身份认证是指从信号中提取反映设备身份的特征，然后生成可用于识别设备的指纹，再通过机器学习算法识别设备指纹，从而实现设备身份的认证。通常设备指纹具有唯一性、可检测性。设备身份认证技术是将信号分析与处理技术与机器学习技术相结合。目前该领域的相关研究成果主要基于暂态信号、调制信号、频谱响应以及传感器响应产生的指纹进行设备身份识别。基于机器学习技术的设备身份认证流程如图 4 所示，包括测量信号、提取信号特征、降低维度、生成指纹及指纹识别 5 个阶段。

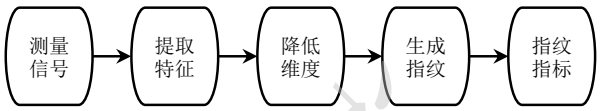


图 4 基于机器学习的设备身份认证流程

Tekbas 等人[66]提出利用设备开关的暂态特征实现设备身份的指纹识别。具体过程为，测量设备暂态信号中含有幅度和相位信息的复包络，将暂态信号的方差变量作为暂态特征，并采用自组织映射网络（Self -Organizing Maps，SOM）降低暂态特征的数据维度，最后利用概率神经网络（Probability Neural Network，PNN）对设备指纹进行识别。在不同的电源电压、环境温度以及信道噪声的条件下实验，发现环境变化让识别性能明显下降，但可通过在环境变量更大差异的条件下收集暂态信号，使性能得到补偿，从而提高识别准确率。

Brik 等人[67]针对无线网络设备，提出了基于调制信号的设备身份认证技术。该方法使用 SVM 和 K 近邻算法分别进行认证设备身份。实验结果表明，SVM 的识别率比 K 近邻算法较高，但运算速度较慢。原因在于，SVM 算法需要将输入数据映射到更高维的向量空间，而 K 近邻算法相对简单，不需要做数据预处理。该方法的缺点是需要收集设备的传输数据，从而带来一定的隐私安全问题。

Danev 等人[52]首次提出了基于调制信号和频谱响应信号的 RFID 设备应答器的物理层认证技术。具体过程为，首先提取设备的调制信号和频谱响应，采用 PCA 降低特征数据的维度，然后生成设备指纹，最后再利用 K 近邻算法实现设备识别。但在交叉验证时，发现该方案仅能在可控的环境下进行，距离较远时（例如超过 1 米）则无法识别设备指纹。

Dey 等人[68]专门研究了利用机器学习实现智能手机和平板电脑设备的身份认证。不同于传统的通过 cookie 和设备 ID 来进行身份认证，文章利用不同的传感器对同样的运动刺激会产生不同的响应的原理，提出利用设备内部传感器指纹实现设备的认证。具体流程为，首先从加速度传感器产生的运动路径的时域、频域信号中提取了若干特征数据，生成传感器指纹；然后利用随机森林算法对设备传感器进行指纹识别。这种方法可以追踪用户踪迹，因此对用户隐私安全造成威胁。

综上，在利用机器学习对暂态信号、调制信号、频谱响应以及传感器响应产生的指纹进行设备身份识别时，外部环境变量会对识别效果产生较大影响，因此今后的研究需要进一步提高信噪比和识别准确率。此外，现有的研究方法主要利用单一或少量的设备身份指纹要素，因此多指纹要素结合的设备身份认证方法也是未来探索的方向之一。同时，现有方法在采集数据时可能都会涉及到用户隐私，因此在未来的研究中要加强用户隐私保护。

3.2.2 物理层边信道攻击

含有密码算法的设备在工作状态时，会在电源功耗消耗、密码算法执行时间、电磁辐射、故障情况的输出等方面产生与密钥相关的变化信息，这些信息即为物理层边信道信息。物理层边信道攻击则是利用物理层边信道信息找出设备的加密信息的一种攻击方式。在物理层边信道攻击研究中，与机器学习结合的攻击方法主要有模板攻击和能耗分析攻击。

在模板攻击研究中，Hospodar 等人[69]将最小二乘支持向量机（Least Squares Support Vector Machine，LS-SVM）算法应用到模板攻击边信道中。在模板刻画阶段，首先使用多元高斯分布刻画边信道泄漏数据的分布特征，并利用皮尔逊相关系数和 PCA 算法得到特征；在密钥恢复阶段，利用 LS-SVM 算法进行模式匹配实施攻击。作者实验发现 LS-SVM 算法对于攻击效果和汉明重量泄漏有显著的影响，但对能耗和时间的影响却不明显。然而，在模板攻击中，边信道攻击通常假设密码设备完全被控，为了对该假设进行松弛，Lerman 等人[70]提出利用半监督学习算法一般化模板攻击，同样能够推出汉明重量子密码。该方法虽然准确度小于普通的模板攻击，但不需要那么强的假设。

模板攻击不仅依赖参数假设、先验知识，而且受限于低维环境。为了解决这个问题，Lerman 等人[71]又提出了基于松弛假设和高维度特征向量的差

分能耗分析（Differential Power Analysis，DPA）攻击，将能耗和密钥关系形式化成一个监督学习任务。特征选取对比实验了排序（Ranking）法、主成分分析法、自组织映射、最大相关最小冗余（mRMR）算法，分类器对比实验了 SOA、SVM、随机森林，然后采用弃一法交叉验证来评估模型的效果，并选择最佳分类器用于推测密钥。实验结果显示，能耗分析攻击的效果优于常规的模板攻击方法。

在物理层边信道攻击研究中，利用物理环境泄露的信息，采用机器学习进行模板攻击、能耗分析攻击是有效的攻击方式。相比而言，由于能耗分析攻击效果优于模板攻击，且不需要很多假设与限制，因此能耗分析攻击是较优的方式。从安全防御角度看，应积极探索应对基于机器学习的物理层边信道攻击的方法，例如增加边信道噪声以混淆边信道信息；此外，还要考虑到基于机器学习的攻击思路能否推广到其他类型的边信道攻击，并提前部署应对措施。

3.2.3 伪基站检测

2G/3G/4G 及兼容模式的基站构成了移动通信网络的基础设施。然而，GSM（2G）网络协议的安全缺陷使得不用经过网络使用者的认证，犯罪分子利用伪基站就可直接攻击用户。例如攻击者利用 ISMI 捕捉器堵塞 3G/4G 网络，迫使用户手机接入 2G 网络，随后向该用户发送垃圾邮件或诈骗短信。

为了检测伪基站，2012 年挪威的安全专家首次提出基于机器学习的 ISMI 捕捉器检测系统[72]，主要包括在线检测和离线学习两部分。在线检测由若干单类 SVM、神经网络等构成的异常检测器组成，主要利用的环境属性特征包括 2G 和 3G 之间的模式转变、真正基站检测到的手机信号消失的时间、加密的禁用等；然后通过集成算法集成强检测器，再结合安全专家判断，最后将综合结果反馈给离线学习部分以更新检测器参数。2017 年 Li 等人[9]利用将近 1 亿的众包数据和 SVM 算法，研制了大规模伪基站检测系统。整个系统由 3 个模块（即内容安全分析，内容分析和 SVM 聚类）和 4 个数据集（权威电话号码表、基站位置数据库、WiFi 位置数据库和短信日志库）组成，完成伪基站定位和短信分类。

在基于机器学习的伪基站检测的研究中，文献[72]存在着没有真实的伪基站训练集、无法展示不同基站设备间的关联关系等问题，仅仅是将机器学习机器检测伪基站作为一个可能的方案，因此需要进一步研究如何收集真实的训练集数据、确定伪基站的区域。Li 等人[9]虽然进行了真实场景应用，解决了训练数据集的问题，并能够定位伪基站，但缺乏真实的伪基站测试集，查全率有待进一步研究。

### 3.3 系统软件安全

在系统软件层面，目前机器学习在系统软件安全中的研究主要集中在漏洞分析与挖掘、恶意代码分析、用户身份认证以及虚拟化安全等方面。

3.3.1 漏洞分析与挖掘

漏洞（Vulnerability）是指系统在硬件、软件及协议的具体实现中或系统安全策略设计上存在的缺陷和不足，从而威胁、损坏单元计算系统的安全。从最早的莫里斯蠕虫到 2017 年 5 月爆发的 WannaCry 勒索病毒，无不是利用系统中的漏洞对系统进行攻击，并利用网络进行传播，因此系统软件中的漏洞识别无疑是网络空间安全研究中的重点。漏洞识别研究已有多年，例如利用漏洞的特征进行识别、随机测试技术（例如模糊测试）以及利用污点分析、符号执行等分析方法。然而，这些方法实际中很难被有效利用，并且只有少部分的安全缺陷能够自动地被识别，大部分的安全漏洞仍然依靠冗长的代码审计。

为了加快漏洞发现和人工审计的过程，Yamaguchi 等人[73]从基于函数形式的源码中提取出 API 符号，利用 PCA 自动地识别 API 用途的特征，从而发现了零日漏洞。在深度的二进制程序漏洞静态分析过程中，加州大学伯克利分校的研究人员提出利用深度学习中的循环神经网络（Recurrent Neural Network，RNN）识别函数[74]。Yamaguchi 等人还提出代码漏洞漏检自动识别方法，即通过语法抽象图和词袋模型（代码的特征描述），利用机器学习方法进行分析[75]。该方法能够准确地识别出是否是真正的漏检，并且还挖掘了若干零日漏洞。

此外，还有部分研究集中在漏洞预测方面。例如 2013 年 Pang 等人[76]提出基于 SVM 集成学习的软件组件早期漏洞的识别方法；Scandariato[77]基于机器学习文本挖掘方法预测软件源码中含有的安全漏洞；2015 年 Pang 等人[78]又利用自然语言处理中的 N-Gram 模型和统计特征选择来预测漏洞；更进一步地，Long 等人[79]提出基于人工手写补丁的特征，利用概率模型自动生成补丁，以自动修复漏洞。

在漏洞分析与挖掘方面，近几年刚开始出现基于机器学习的漏洞识别、漏洞预测和漏洞修

复，还没有形成一个较为成熟的应用体系。更为重要的是，根据莱斯定理（Rice's Theorem），利用一个程序自动地检测另一个程序中是否含有漏洞，在一般情况下是不可判定的[75]。在这个限制下，现有的漏洞挖掘研究主要集中在发现特定类型的漏洞，因此利用机器学习推断未知漏洞在理论上是否可信仍是一个问题。总之，未来研究除了考虑进一步提升机器学习在漏洞挖掘的应用性能（例如识别率、查全率），也应考虑到机器学习应用在漏洞挖掘领域的理论支撑。

3.3.2 恶意代码分析

恶意代码通常指具有恶意功能的应用程序，包括木马、蠕虫、病毒等。恶意代码分析通常分为静态分析和动态分析。静态分析通过分析程序的指令与结构来确定是否具有恶意功能；而动态分析是在隔离环境（例如模拟器、沙盒）运行状态下，综合分析运行行为，从而确定是否具有恶意功能。目前，已有大量研究利用机器学习技术分析代码量庞大、代码特征或运行行为特征复杂的恶意软件。

在恶意代码静态分析方面，Arp 等人[26]收集软件尽可能多的特征，并嵌入联合特征向量空间中，通过 SVM 检测其中的恶意代码。Nath 等人[80]使用 N-Gram 模型从原始恶意代码中提取训练集，然后对比了朴素贝叶斯、SVM、决策树等算法，并将这些弱分类器算法集成提升为强分类器算法，即 AdaBoost 算法，取得了良好的检测效果。

在恶意代码动态分析方面，为了定期地更新可疑文件的特征及减少人工分析时间，Nissim 等人[81]提出了基于主动学习框架的捕获新型恶意软件的方法。Rootkit 恶意软件能够在内核中隐藏自身及指定的文件、进程以及网络链接等信息，对操作系统安全威胁极大。为此 Wilhelm 等人[82]通过二进制文本和运行行为提取特征，利用朴素贝叶斯分类算法对特征进行分类，判断内核驱动是否含有 Rootkit。近几年移动终端恶意软件问题凸显，大量研究借助机器学习动态分析移动终端恶意软件，例如 Narudin 等[49]选取了信息、内容、时间和连接四个网络行为特征，使用贝叶斯网络和随机森林方法取得了 99.97%的准确率。Chen 等人[83]基于动态行为、请求许可、请求时间序列、敏感程序接口四个特征，提出了基于机器学习的恶意软件检测流处理框架，并在分布式实时计算框架 Storm 之上实现了系统原型，同时提高了检测效率和准确率。

随着移动应用的广泛应用，移动端的恶意软件检测逐渐成为了研究的热点。此外，随着恶意软件的不断升级，仅采用静态分析技术无法取得较好的检测效果，恶意代码的动态分析是该方向的发展趋势。

3.3.3 用户身份认证

与 3.2.1 节中设备身份认证不同，用户身份认证研究用户与硬件设备或用户与系统之间的认证。在基于机器学习的用户身份认证研究方面，主要有利用机器学习攻击传统用户身份认证方法和利用机器学习设计新的用户身份认证机制两个研究点。

在利用机器学习攻击传统用户身份认证方面，针对设备用户认证过程中的验证码，斯坦福大学的 Golle 等人[84]最早提出了基于机器学习的验证码攻击，即程序自动识别验证码。该验证码的分类器是多个 SVM 分类器的结合，用以训练从验证码图片中提取的颜色和文本特征。Yue 等人[85]基于计算机视觉，首次提出了自动盲识别触屏设备的输入密码，在其攻击模型中，通过 DPM 模型来检测和追踪目标设备，并利用光学流算法自动识别触摸帧以及利用 K-Means 聚类算法来识别触摸点。Liu 等人[86]基于可穿戴设备传感器边信道信号，推断出了用户的键盘密码输入，其技术思路为：首先建立可穿戴设备运动模型，然后从可穿戴设备提取数据、预处理、特征提取，接着采用经典的机器学习算法（例如随机森林、K 近邻、SVM、神经网络等），对比试验结果显示，K 近邻算法最佳。实验对带着智能手表的志愿者进行，在数字键盘输入时推断出了他们的银行 PIN 码，在字符键盘输入时推断出了他们的 POS 机密码的英文输入。

在基于机器学习和生物特征的用户身份认证设计方面，Zheng 等人[87]基于触屏特征——加速度、压力、大小和时间，提出了一个免打扰的用户认证机制，实验收集了 80 个用户的触屏数据，采用单类学习算法来验证是否是合法用户，结果显示该方法的等错率（Equal Error Rates，ERR）为 3.65%。Giuffrida 等人[88]基于传感器增强的击键行为特征，在安卓操作系统上实现了包含多种特征提取和识别算法（包括 SVM、朴素贝叶斯、马氏距离算法、K 近邻等）的用户认证系统；实验发现当识别算法为 K 近邻且 K=1 时效果最佳，且比基于动作的认证机制高了一个数量级的等错率，比传统的击键方案准确率高了两个数量级的等错率。Deng 等人[89]基于斯坦福大学公开的智能手机击键认证数据集，利用 DNN 来提高击键认证技术。更进一步地，

Kobojek 等人[90]将基于 DNN 改进的 LSTM 和 GRU 网络结构应用到击键认证设计中。此外，Li 等人[91]为了抑制智能手机非授权使用，首次提出基于手势特征的再次认证身份系统，并采用 SVM 来识别手机拥有者滑动手机屏幕的手势特征。

**表 3 机器学习在用户身份认证技术中的应用**

| 安全问题 | 安全特征 | 机器学习算法 | 相关文献 |
|---|---|---|---|
| 用户身份认证攻击 | 验证码的颜色和文本特征 | SVM | [85] |
| | 计算机视觉、触摸点 | K-Means 算法 | [86] |
| | 边信道信号 | 随机森林、K 近邻、SVM、神经网络 | [87] |
| 用户身份认证设计 | 触屏特征（加速度、压力、大小和时间） | 单类学习算法 | [88] |
| | 击键行为 | SVM、朴素贝叶斯、K 近邻 | [89] |
| | 击键 | DNN | [90] |
| | 击键 | LSTM | [91] |
| | 手势 | SVM | [92] |

机器学习在用户身份认证中的相关研究如表 3 所示，一方面利用机器学习对传统用户身份认证方式进行攻击的手段层出不穷，另一方面利用机器学习技术构建用户身份认证机制，提供更强的安全性能，但同时也存在一些问题，例如认证准确率不高、训练数据涉及用户隐私等。因此，在今后的用户身份认证研究中，不仅需要重点关注如何应对基于机器学习的用户身份认证攻击，还要加强基于机器学习的用户身份认证机制的隐私保护。

3.3.4 虚拟化安全

以虚拟化技术为基础的云计算在网络空间中已经得到广泛应用。目前虚拟化安全问题主要是虚拟机隔离安全，而边信道攻击又是虚拟机隔离安全中的一大威胁[92]。本节将介绍基于机器学习的虚拟机边信道攻击及虚拟机环境恶意行为检测。

在基于机器学习的虚拟机攻击研究中，共享高速缓存架构例如一级高速缓存（Level 1 Cache）、最后一级高速缓存（Last Level Cache，LLC）常被作为跨虚拟机攻击的通道。2012 年 Zhang 等人[93]首次提出了通道驱动的跨虚拟机的边信道攻击。在对称多处理（Symmetric Multi-Processing，SMP）环境下，将一级高速缓存作为攻击通道，通过素数探针实现同一物理机上的恶意虚拟机从受害虚拟机中提取细粒度信息，再利用 SVM 算法分类 Cache 信息，并基于代码路径的隐马尔可夫模型来提高 SVM 输出的准确性，从而最终获取了受害虚拟机用开源加密库 Libgcryp 实现的 EIGamal 加密的秘钥。2017 年 Gulmezoglu 等人[94]也利用 SVM 算法和高速缓存通道进行跨虚拟机攻击，对来自 Phoronix 网站的 40 个基准测试应用进行测试，实验结果表明，一级高速缓存器取得了 98%的分类准确率，最后一级高速缓存器取得了 78%的分类准确率；而在夹带噪声的 Amazon EC2 跨虚拟机环境下，25 个基准测试应用的准确率下降到了 60%。

在虚拟机环境恶意行为检测研究中，Fischer 等人[95]在云数据中心虚拟机恶意软件防护架构中，设计了一个基于机器学习算法的数据分析模型，以此作为云数据中心虚拟机入侵检测器，当检测到恶意行为时立即启动决策引擎响应，如隔离或重启虚拟机或者启动虚拟机网络管理重新配置、分配虚拟机资源。

总体而言，利用机器学习技术研究虚拟化安全的工作目前还不多。现有的基于机器学习的跨虚拟机攻击虽然存在基于同一物理机的前提要求，但利用机器学习技术分析云计算环境信息、攻击虚拟机，例如 Zhang[93]、Gulmezoglu[94]分别密钥推测、应用程序类型推测的角度进行攻击，这是云计算中面临的巨大威胁。因此需要特别关注如何应对基于机器学习的虚拟机攻击，并有针对性的部署防御措施。

### 3.4 小结

本节从芯片、系统硬件及物理环境及系统软件三个层面，详细介绍了已有的机器学习在系统安全中的研究工作，其主要的安全特征、机器学习算法以及相关文献总结如表 4 所示。

总体上，已有部分研究人员利用机器学习的优势特点对系统安全进行了大量尝试与研究，但还存在着以下问题及未来可能的研究方向：（1）在芯片安全层面：需要进一步提高劣质芯片和硬件木马的检测精度，以及增强 PUF 的抗机器学习攻击的能力。（2）在系统硬件及物理环境安全层面：在利用各种信号进行设备身份认证时，需要在噪声大的情况下提高识别准确率，还要注意隐私保护的问题。针对基于机器学习的物理层边信道攻击，可以考虑增加边信道信号噪声以混淆、对抗攻击。伪基站检测方面，未来需要进一步解决伪基站数据集收集、伪基站定位精度、查全率三大问题。（3）在系统软

件层面：针对漏洞分析与挖掘，虽然利用机器学习可以推断未知漏洞，但可信度还有待研究。对于恶意代码分析，随着恶意软件的不断升级，有从静态分析到动态分析的趋势，并利用不断发展的机器学习算法进一步提高恶意软件检测准确率。在利用机器学习攻击传统的用户身份认证机制时，可以考虑增加视觉混淆、边信道噪声等，以增强认证机制的安全性；利用机器学习设计新的用户身份认证机制时，需要进一步提高识别准确率以及对训练数据进行隐私保护。对于虚拟化安全，需要考虑应对基于机器学习的虚拟机边信道攻击以及构建虚拟机入侵检测器。

**表 4 机器学习在系统安全中的应用**

| 系统安全 | 安全问题 | 问题抽象 | 主要特征 | 机器学习方法 | 参考文献 |
|---|---|---|---|---|---|
| 芯片安全 | 劣质芯片检测 | 分类 | 芯片外形、边信道信号 | OC-SVM、PCA、SOA、ANN | [56-58] |
| | 硬件木马检测 | 分类/聚类/降维 | 边信道信号、芯片原理图 | OC-SVM、PCA、非线性回归、K 近邻、SVM | [38,48,61-62] |
| | PUF 攻击 | 分类 | PUF 激励-响应对 | 逻辑回归、SVM、进化策略、ANN | [64-65] |
| 系统硬件及物理环境安全 | 设备身份认证 | 聚类/降维 | 暂态信号、调制信号、频谱响应、内部传感器响应 | SOA、PNN、SVM、K 近邻、PCA、随机森林 | [45,66-68] |
| | 物理层边信道攻击 | 分类/聚类 | 边信道信号分布特征、差分能耗 | LS-SVM、学习排序法、PCA、自组织映射、mRMR 算法、SOA、SVM、随机森林 | [69-71] |
| | 伪基站检测 | 分类 | 2G 和 3G 之间的模式转变、真正基站检测到的手机信号消失的时间、加密的禁用、基站位置数据库、WiFi 位置数据库和短信日志库 | 神经网络、SVM | [9,72] |
| 系统软件安全 | 漏洞分析与挖掘 | 分类/聚类 | 源码 API 用途、代码语法 | PCA、RNN、SVM、集成学习、N-Gram 模型 | [73-79] |
| | 恶意代码分析 | 分类/聚类 | 二进制文本、运行行为，信息、内容、时间和连接，动态行为、请求许可、请求时间序列、敏感程序接口 | SVM、AdaBoost、贝叶斯网络、随机森林、K 近邻 | [26,49,80-83] |
| | 用户身份认证 | 分类/聚类 | 触屏特征、击键行为、验证码颜色与文本、计算机视觉、传感器边信道信号 | SVM、朴素贝叶斯、马氏距离算法、K 近邻、DNN、LSTM 和 GRU 网络、K-means | [84-91] |
| | 虚拟化安全 | 分类 | 虚拟机边信道信息 | SVM | [93-95] |

## 4 机器学习在网络安全研究中的应用

网络安全是网络空间安全中的重要支柱，网络基础设施的安全为互联网的可靠运行的提供了基础，各项网络安全检测措施为互联网的各项活动的开展提供了安全的通信保障。本节主要介绍机器学习在网络基础设施的安全以及网络安全检测方面的相关研究成果，包括机器学习技术在 BGP 的异常检测、恶意域名检测、僵尸网络检测、网络入侵检测以及恶意加密流量的识别中的应用研究。

### 4.1 网络基础设施安全

路由系统和域名系统是网络空间中重要的网络基础设施，这些系统安全可靠的运行是各项网络活动安全开展的基础和前提，因此路由安全和域名系统的安全一直是学术界和工业界关注的重点。近年来，采用机器学习进行 BGP 的异常检测以及 DNS 的恶意攻击检测，均取得了一定的检测效果。

#### 4.1.1 BGP 的异常检测

边界网关协议（Border Gateway Protocol，BGP）是互联网的核心路由协议，互联网的域间路由通过 BGP 路由信息交换来完成。但 BGP 缺乏一个安全可信的路由认证机制，无法对邻居自治系统

（Autonomous System，AS）宣告的路由信息进行完整性和真实性的验证。这一缺陷导致路由器面临多种攻击，其中前缀劫持（Prefix Hijacking）、异常 BGP 的更新消息等问题严重影响了互联网的连通性和安全性[96]。早期的异常路由识别方法采用诸如统计分析方法、信号处理技术等技术分析流量行为模式，这些方法难以计量所有可能的异常路由的分布及维度[97]。异常路由的检测是通过提取当前 BGP 更新消息的特征或时序特征，将当前流量识别为正常路由或异常路由，该问题可以抽象为机器学习二分类问题。因此，研究者尝试采用 SVM、隐马尔科夫模型、决策树、朴素贝叶斯、LSTM 等算法进行异常路由检测及前缀劫持定位研究。

在异常路由检测研究中，许多研究人员采用了在公开数据集 RIPE RIS[26]和 Route Views[27]，从中提取不同的特征，然后利用不同的机器学习技术进行研究。例如，Li 等人[21]对公开路由数据集中提供的 BGP 更新消息提取了 37 个特征，包含 BGP 通告的数量、平均 AS-PATH 的长度、最长 AS-PATH 的长度、IGP 包的数量、EGP 包的数量包等，然后采用决策树和柔性粗糙集技术对上述 37 个特征分别进行特征选择，去除因数值化而产生的重复特征，由此分别生成特征个数不同的 3 个数据集。随后采用决策树和 ELM（Extreme Learning Machine）算法[98]两种算法分别对不同的数据集构建分类器识别 BGP 的异常路由，取得了平均 80.08%的准确率。该方法不需要修改协议，风险较小。Al-Rousan 等人[99]利用 SVM 算法和隐马尔科夫模型对 BGP 更新消息进行特征选择，对异常 BGP 的检测取得了 81.5%的准确率。然而，上述方法均是基于短时特征的检测方法，并且未考虑流量的时间序列特性以及流量的随机性，因此无法对具有噪音且动态变化的突发流量进行检测。为此，Cheng 等人[34]在对 BGP 流量的时间序列分析的基础上，针对网络流量的多维度的时间特性以及在一个滑动窗口中的流量的历史特征，选取了 33 个具有时间序列特性的流量特征，提出了采用长短期记忆网络（Long Short-Term，LSTM）模型对 BGP 异常路由检测，比 SVM[99]、朴素贝叶斯[100]及 AdaBoost[101]方法提高了 10%的识别率。

针对前缀劫持定位问题，Qiu 等人[102]提出了一个轻量级且能增量部署的 LOCK 方案，该方案利用层次聚类算法（Hierarchical Clustering）对 LOCK 系统中部署的大量监控器分成若干簇，每个簇中的监控器到目标前缀具有相似的路径。当前缀被劫持时，每个簇中的监控器基于目标前缀被污染的路径的概率进行排名，排名最高的监控器最有可能定位到前缀劫持，选择每个簇中的排名最高的监控器监控目标前缀，由此提供精准的前缀劫持定位。

总体而言，异常路由检测从 BGP 更新消息或时序中提取异常信息对异常路由报警，前缀劫持行为定位以完善 BGP 安全为目标，对前缀劫持行为进行定位并报警。目前机器学习技术应用于异常路由检测，检测准确率仍不高，且存在误报和漏报的可能性；同时，该方法仅局限于模型的构建与探讨，未在实际网络中部署实施。因此在今后的研究中需要对异常 BGP 的特征作深入研究，采用新兴的机器学习技术或机器学习集成方法提高 BGP 异常检测的准确率以及误报率，进一步探索在实际网络环境中如何实施机器学习技术应用于异常路由检测中。

#### 4.1.2 恶意域名检测

域名系统是互联网中的核心应用之一，域名系统经常成为攻击的目标，或被攻击者利用作为攻击工具，因此域名系统的安全一直是网络安全的研究热点。早期的恶意域名检测方式是在域名系统、防火墙或网络入侵检测系统中设置恶意域名黑名单或拦截名单，该方法极易被攻击者躲避检测。随后出现的基于查询请求数的方法[103] [104]，存在误报率较高且无法检测未知异常域名检测等为问题。近年来，应用机器学习技术构建恶意域名的检测规则是该领域新的研究方向。

基于机器学习的恶意域名检测通常为离线模型和在线模型相结合[105]，一般流程如图 5 所示。离线模型中，将带有标签的合法域名和恶意域名作为训练数据集，从中提取基于网络层的特征、基于区域的特征、基于时间的特征、基于 DNS 应答的特征、基于 TTL 的特征或者基于域名信息等不同层次的特征，然后选取决策树、X-Means 聚类算法等构建训练模型，同时可采用诸如 malwareurl.com①、McAffee Site Advisor②或 Norton Safe Web③等网站提供的已知的域名数据集对训练模型进行验证和参数调整。在线检测模型中，网络

① http://www.malwareurl.com/

② http://www.siteadvisor.com/

③ https://safeweb.norton.com/

中实时采集的域名流量经过被动域名查询分析，进行域名特征提取，如果是已知的域名信息，即已标记的域名特征，则输入训练模型继续训练；如果是未知的域名信息，即无标签的特征，则输入已训练好的分类器对该域名进行判别是否为恶意域名。

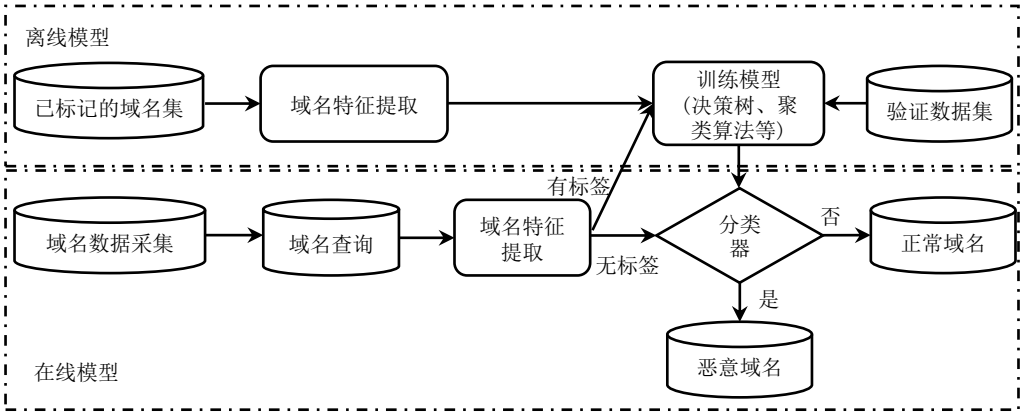


图 5 基于机器学习的恶意域名检测流程

例如，动态域名信誉系统 Notos[46]使用已知垃圾邮件中的恶意域名和 Alexa.com 网站中已知的合法域名作为被标记的域名数据，从中提取基于网络特征、基于区域特征以及基于证据的特征属性，使用 X-Means 聚类算法先后对网络特征及区域行为特征进行聚类，然后利用聚类后的特征向量与基于证据特征向量使用 LAD（Logit-boost Stategy）决策树对新域名的进行信誉评分，评分采用{0,1}表示，0 表示恶意域名，1 表示正常域名。该模型的建立主要基于网络特征，对于新地址空间映射的恶意域名无法进行检测。域名分析系统 Exposure[106]从多个来源的已知的域名数据中提取了短期域名、域名的访问比例、不同 IP 地址的个数、不同国家的数量、共享 IP 的数量、不同 TTL 值的数量、域名的长度等 15 个行为特征，采用 J4.8 算法（C4.5 算法在开源机器学习软件 Weka①中的实现）构建决策树。Exposure 系统与 Notos 系统相比，Exposure 系统不依赖于历史恶意数据进行建模，并且模型训练过程中使用的训练数据和训练时间较少、限制条件也较少。

机器学习技术除了用于恶意域名的检测外，还用于随机域名的生成。在 2016 年的 BSidesLV② 会议中，基于深度学习的 DGA 算法 DeepDGA 被提出[19]，该方法采用 Alexa 网站上收录的知名域名作为训练数据，利用 LSTM 算法和生成对抗网络构建模型，生成的域名与正常网站域名无异，针对机器学习技术生成的随机域名较难检测出，易于被攻击者利用。

采用机器学习技术检测恶意域名虽然取得了一定的成效，但仍然存在两方面的问题，一是如果攻击者了解域名检测系统的原理，就可以轻易的逃避机器学习构建的检测系统；因此未来的研究工作中可以采用诸如生成对抗网络等机器学习技术构建新型的恶意域名检测系统。二是现有的系统主要取决于已知的恶意域名作为训练数据构建的检测系统，对于例如采用深度学习技术生成的随机恶意域名检测效果不佳，因此未来的研究中需要进一步发挥机器学习技术在未知恶意域名的检测中的作用。

### 4.2 网络安全检测

网络安全检测主要指对网络的安全状态或者面临的风险进行分析和检测，对不同接入网络的行为进行分析和控制，以发现潜在的威胁或正在进行的攻击。本节具体分析了机器学习技术在僵尸网络的检测、网络入侵检测技术以及恶意加密流量的识别中的研究现状。

#### 4.2.1 僵尸网络的检测

在僵尸网络中存在大量被僵尸程序感染的主机，通常称为 bot 或 Zombie，受攻击者的控制进行恶意网络攻击，例如分布式拒绝服务（DDoS）、垃圾邮件（Spam）、网络钓鱼（Phishing）以及窃取信息等恶意活动。僵尸网络的活动主要分为传播、命令与控制（C&C）、攻击三个阶段，其中命令与控制是僵尸网络核心工作机制[107]。由于不同的僵尸网络的传播、命令与控制以及攻击方式各不相同，传统的僵尸网络检测采用效率极低的人工分析的方式[108]。而机器学习的自学习能力为提

① http://weka.wikispaces.com/

② BSidesLV 是一个非营利组织，每年为安全工程师或安全公司提供为期两天的开放论坛。

高僵尸网络的检测效率提供了可能性。将机器学习应用于僵尸网络检测中，首先从骨干网或者企业网中的流量以及日志信息中提取流量特征或者行为特征，然后利用 X-Means、随机游走等聚类算法、SVM、随机森林及关联规则等算法实现检测。根据检测特征的不同，应用机器学习技术的僵尸网络的检测分为基于网络流量分析和基于关联分析的检测技术。

从网络流量角度分析，僵尸网络的通信行为从流量角度分析具有一定的关联性和群体相似性，因此许多研究者通过聚类方法分析不同的网络流量特征，从而检测僵尸网络。例如，Nagaraja 等人提出的 BotGrep[109]是一种通过对网络流量行为分析检测结构化 P2P 僵尸网络的方法，该方法通过从骨干网中采集的流量提取结构化 P2P 网络的主要特征，例如快速收敛时间，然后利用随机游走(random walk)聚类算法构造结构化 P2P 网络的子图，再结合蜜罐等检测技术来判断是否为僵尸网络。Antonakakis 等人提出的 Pleiades 系统[110]采用图聚类和隐马尔可夫模型检测基于 DGA 的僵尸主机，不仅能够识别已知的基于 DGA 的恶意软件族，而且发现了新的基于 DGA 的僵尸主机。Zhang 等人[111]利用网络中采集的 bot 查询流量，第一次分析了 bot 查询意图以及攻击模式，采用基于主题的单链层次聚类算法检测 bot 在查询引擎中的查询是否具有攻击行为。Jacob 等人提出的 Jackstraws 系统[112]利用主机信息，采用图聚类方法自动捕获不同类型的 C&C 活动，实现了从僵尸流量中发现 C&C 连接。此外，其他的机器学习方法也用于尝试检测僵尸网络。例如，Bilge 等人[113]采用随机森林算法对 Netflow 采集的网络流进行分析，进而对僵尸网络进行检测。Chen 等人[114]采用能够自适应特征集和训练分类标签的 LS-SVM 算法，对流分析抽取图特征用以检测僵尸网络。

除了提取僵尸网络流量的流量特征外，僵尸网络的通信行为往往和一些恶意事件在时间或空间上有关联。因此，许多研究工作采用流量特征与日志信息的关联分析检测技术集宁僵尸网络检测。例如，BotSniffer 检测系统[115]对同一局域网中 bot 活动的时间和空间关联性进行分析，从 IRC（Internet Relay Chat）和 HTTP 流量中识别出可疑的僵尸网络 C&C 连接，然后结合异常事件日志，采用 X-Means 聚类方法进行关联分析来进行检测集中式的僵尸网络。BotMiner 检测系统[116]则不依赖于具体协议以及僵尸网络结构，将所有的流量按目的地址和端口进行聚合，利用 X-Means 聚类后的流量属性以及异常事件的日志进行关联分析来检测僵尸网络。RB-Seeker[117]系统采用了适合小样本数据集的线性 SVM 算法和关联分析技术相结合检测重定向僵尸网络。该方法首先将网络流量数据、spam 信息以及 DNS 日志进行关联分析，利用线性 SVM 技术检测出恶意域名，然后与 DNS 日志进行关联找出重定向的僵尸网络。此外，对于加密流量的检测，不依赖于深度包检测技术而使用关联规则技术[118]可以识别 C&C 通信中的关键特征，从而自动识别受感染的主机。

综上所述，在僵尸网络检测研究中，表 5 总结了当前研究成果所采用的机器学习技术及选取的检测特征，现有研究成果表明，由于网络流量数据量巨大以及僵尸网络通信行为复杂，使得僵尸网络流量的特征选取较困难，现有研究多数针对特定类型的僵尸网络，且通过聚类方式的流量分析技术检测精度很难达到实际应用需求。而将网络流量行为和恶意事件日志进行关联分析的检测技术，虽然在一定程度上提高了检测精度，也仍有其局限性。例如，BotSniffer 采用固定的通信协议以及特定类型的僵尸网络结构，BotMiner 虽然不依赖于固定的通信协议和僵尸网络拓扑结构，但采用了计算量较大的 X-Means 聚类算法，从而检测时间较长。RB-Seeker 系统只针对特定的僵尸网络活动，检测系统的可扩展性差；BotFinder 系统虽然可以检测多种僵尸网络活动，但只对其中的三种僵尸网络类型检测效果较好。因此随着僵尸网络传播、控制及攻击形式的多样化，如何对僵尸网络流量进行有效的特征提取、如何能够全面提高各种类型僵尸检测效率及精度，都是需要进一步深入研究的方向。

4.2.2 网络入侵检测技术

不网络入侵检测的概念于 20 世纪 80 年代提出，网络入侵检测即根据网络流量数据或主机数据来判断系统的正常行为或异常行为，可以抽象为分类问题。利用机器学习在解决分类问题的强大能力，研究人员对基于机器学习的网络入侵检测技术作了大量研究[6,7,9]。经过多年的发展，网络入侵检测技术趋于稳定，近五年机器学习与入侵检测相结合的新研究成果相对较少。本文仅取

了有代表性的部分研究成果阐述机器学习技术在网络入侵检测中的应用现状。

根据入侵检测系统中检测引擎使用方法的不同，网络入侵检测分为误用检测、异常检测及混合检测。误用检测是以已知攻击为特征，将入侵行为与正常行为按照已知的特征区分开来。该类方法效率高且误报率低，但只能发现已知的入侵、漏报率较高，并且特征的维护多采用人工方式完成。异常检测指将当前网络行为与系统正常行为模式进行比较，若两者偏差较大，超过了预定义的阈值，则认为系统出现了异常或被入侵。异常检测对新的攻击类型敏感，能够有效发现新的攻击，并且能够检测零日漏洞。但是系统正常行为模式的学习通常比较复杂，由此建立的异常检测系统对未知攻击可能会产生较高的误报率。混合入侵检测是指将误用检测与异常检测相结合，用于提高已知入侵检测率并降低未知攻击的误报率。从已有研究成果看，大多数检测方法都是异常检测与混合入侵检测一起使用，鲜见文献单独使用异常检测或混合检测技术的[130]。经典的机器学习算法均可用于网络入侵检测模型中，因此本文按照经典的机器学习算法分别介绍在网络入侵检测技术的研究现状。

**表 5　机器学习在僵尸网络检测中的应用**

| 检测技术 | 特征 | 机器学习算法 | 精度 | 误报率 | 相关文献 |
|---|---|---|---|---|---|
| 基于网络流量分析的检测技术 | P2P 流量 | 随机游走 | 93~99% | 0.77%~0.92% | [109] |
| | DGA 流量 | 图聚类、隐马尔科夫模型 | 95%~100% | <1.4% | [110] |
| | Bot 意图查询 | 单链层次聚类算法 | 94.13%- | - | [111] |
| | 主机信息 | 图聚类 | 81.6% | 0.2% | [112] |
| | NetFlow | 随机森林 | 30%~70% | 0.5% | [113] |
| | 图特征 | LS-SVM | 98% | - | [114] |
| 基于关联分析的检测技术 | C&C 异常命令结合异常日志信息 | X-Means 聚类、关联规则 | 100% | <6% | [115] |
| | 流量的目的地址和端口聚合集合结合异常日志信息 | X-Means 聚类、关联规则 | 75%~100% | <0.0091% | [116] |
| | 网络流量数据关联 spam 信息以及 DNS 的日志记录 | SVM、关联规则 | 96.7%~99.3% | <0.008% | [117] |
| | C&C 通信中的关键特征 | 关联规则 | 49%~100% | 1%~2% | [118] |

将机器学习技术应用于网络入侵检测系统的一般流程如图 6 所示：首先通过网络流量采集工具 Snort、Wireshark 等自行采集数据集或者采用公开的网络入侵数据集，接着对数据集进行预处理从中提取网络入侵特征，然后选择合适的机器学习算法构建入侵检测分类器，对待测数据识别是正常行为还是异常行为。

神经网络是网络入侵检测技术中最流行的机器学习算法。在误用检测研究中，Cannady 等人[27]在数据采集阶段利用 RealSecure$^{TM}$ 网络监控器收集了一万个事件，其中包括利用 Internet Scanner① 和 Satan[119]程序模拟产生的 3000 次攻击事件，在

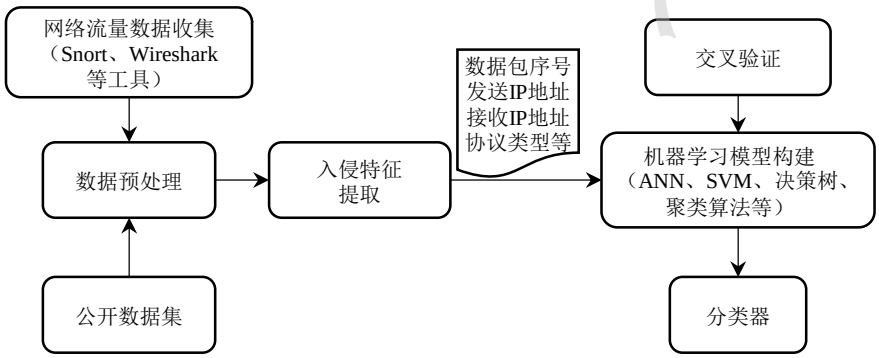


图 6 基于机器学习的网络入侵检测流程

① http://www.iss.net

数据预处理阶段选择了协议标识符（ID）、源端口、目的端口、源地址、目的地址、ICMP 类型、ICMP 代码、原始数据长度和原始数据九个特征，采用神经网络构建多类别分类器，在模型验证阶段利用随机采样方法选择 90%的数据作为训练集和验证集，使用剩余 10%的数据作为测试集。实验结果表明该模型取得了 93%的精度。神经网络构建的误用检测系统易陷入局部最优解并且计算时间较长。在异常及混合检测研究中，Lippmann 等人[13]使用没有隐藏单元的多层感知器组成的神经网络构建异常检测系统，首先将 telnet 会话中关键字（事先预定义列表）出现次数作为神经网络的输入，输出结果为攻击后验概率的估计。随后再将被标记为攻击的实例作为训练数据，利用神经网络生成检测模型，该系统仅实现了 80%的检测率。Bivens 等人[12]使用 DARPA 1999 中的 TCP / IP 数据构建了基于多层感知神经网络的异常检测系统，与分析每个数据包的研究[27]不同，该系统使用时间窗口对多个数据包为一组进行检测，因此该系统能够检测持续时间较长的攻击类型。实验结果表明对正常行为识别率达 100%，但对新攻击的检测效果不佳，误识率（False Acceptance Rate，FAR）达 76%。

关联规则方法用于提取与攻击者入侵行为相关的关联特征。在误用检测研究中，Brahmi[14]则使用 DARPA 1998 数据集，利用关联规则捕获 TCP/IP 参数和攻击类型之间的关系。对四种攻击类型 DoS、Probe 或 Scan、U2R、R2L 的识别率分别达到 99%、95%、75%、87%。当关联规则的维度增大，会带来巨大的计算资源的消耗。在异常及混合检测研究中，Tajbakhsh 等人[50]使用 KDD 1999 数据集中的 30 万个实例，利用模糊关联规则挖掘关联信息，识别率达到 80.6%，FPR 为 2.95%。Apiletti 等人[120]通过从流量数据捕获关联规则，与在线流同时分析实现异常检测。Luo 等人[121]尝试将模糊逻辑应用于关联规则与频率集合中，挖掘数据中模糊序列发生的频率。

贝叶斯分类器对给定序列进行概率计算，例如对系统捕获的一系列异常事件进行概率计算，根据该概率值判断该事件是否为攻击行为。在误用检测研究中，Panda 等人[122]采用了 KDD 1999 数据集，利用开源机器学习软件 Weka 中的朴素贝叶斯分类器[123]，对四种攻击类型 DoS、Probe 或 Scan、U2R、R2L 的识别率达 96%、99%、90%和 90%。该方法与采用神经网络的误用检测模型进行性能对比，取得了较高的识别准确率，但漏报率较高。在异常及混合检测研究中，Kruegel 等人[124]使用 DARPA 1999 数据集，利用贝叶斯网络设计实现入侵检测系统，分析操作系统的调用，并检测运行在 Linux 和 Solaris 系统中的守护进程和 setuid 程序攻击事件。Amor 等人[51]使用朴素贝叶斯分类器实现异常检测系统，该系统使用了 KDD 1999 数据集，实验结果表明没有产生误报，并且正常和异常类的识别率分别达 98％和 89％。Berral 等人[125]在网络中间节点收集本地流量信息，与邻近节点共享信息，网络中的每个节点都采用朴素贝叶斯分类器分类流量，最后各个节点信息聚合，从而提高全网检测 DDoS 攻击的能力。但是如何准确获知网络的流量大小和流量状况分布是在网络层进行 DDoS 检测的难点。在应用层 DDoS 攻击检测研究中，Yan 等人[126]首先分析企业级场景下 DDoS 攻击与防御的博弈策略，提出了利用贝叶斯网络来推断系统可能的状态，用随机变量来描述 DDoS 攻击和防御场景中的系统状态。然后使用贝叶斯网络对多级推理进行建模，对复杂的 DDoS 攻击和防御场景进行评估。该方法仅使用了单个贝叶斯网络，只能对特定的 DDoS 攻击和防御场景进行评估，适用面较窄。

隐马尔可夫模型擅长刻画系统的动态运行特征，也被广泛的应用于网络入侵检测系统中。在误用检测研究中，Ariu 等人[127]为了应对 Web 应用程序攻击，例如 XSS 攻击或 SQL 注入攻击，利用隐马尔可夫模型提取攻击特征，通过检测 HTTP 有效载荷实现多种入侵攻击分类。该方法虽然取了较好的识别准确率，但如何根据已知的 HTTP 有效载荷序列推断隐马尔可夫模型中的隐藏状态是该方法中的难点。在异常及混合检测研究中，Joshi 等人[128]在入侵检测系统中采用 KDD 1999 数据集中的 TCP session 网络流量，选择 41 个特征中的 5 个特征，利用隐马尔可夫模型建模。实验结果表明对未知攻击取得了 79%的识别率。

SVM 以其高效性、稳定性及良好的泛化能力在网络入侵检测系统中取得了较好的识别率。在误用检测研究中，Li 等人[15]利用以径向基函数（RBF）为内核的 SVM 分类器，设计了跟踪特征删除策略和唯一特征选择策略来选择属性子集，将 KDD 1999 数据集的 41 个特征减少到 19 个关键特征，通过 10 倍的交叉验证优化模型取得了

98.62%的识别准确率。在异常及混合检测研究中，Hu 等[129]使用鲁棒支持向量机（RSVM）构建异常检测分类器，在 DARPA 1998 数据集中有噪声的情况下（例如训练数据中存在一些错误标签）进行测试，分类性能良好。Wagner 等人[130]使用 Flame 工具[131]采集了真实攻击数据，包含 NetBIOS 扫描、DoS 攻击、POP 垃圾邮件、安全 Shell、SSH 扫描等数据，再利用 One-Class SVM 分类器设计实现了异常检测系统，实验表明不同类型的攻击识别正确率在 89%~94%之间，FPR 为 0%~3%之间。

**表 6 机器学习在网络入侵检测中的应用**

| 检测类型 | 机器学习算法 | 数据集 | 文献 |
|---|---|---|---|
| 误用检测 | 关联规则 | DARPA1998 | [14] |
| | SVM | KDD1999 | [15] |
| | 神经网络 | 网络层数据包 | [27] |
| | 朴素贝叶斯 | KDD1999 | [122] |
| | 隐马尔可夫模型 | DARPA1998、采集 HTTP 流量 | [127] |
| | 决策树 | DARPA1999 | [132] |
| | 贝叶斯网络 | KDD1999 | [136] |
| 异常及混合检测 | 关联规则 | KDD1999 | [5] |
| | 神经网络 | DARPA1998/1999 | [12,13] |
| | 朴素贝叶斯 | KDD1999 | [51] |
| | 贝叶斯网络 | DARPA1999 | [124] |
| | 朴素贝叶斯 | HTTP 流量 | [125] |
| | 贝叶斯网络 | HTTP 流量或 DNS 流量 | [126] |
| | 隐马尔可夫模型 | KDD1999 | [128] |
| | SVM | DARPA1998 | [129] |
| | SVM | NetFlow 采集流量 | [130] |
| | 决策树 | KDD1999 | [134] |
| | DBSCAN | KDD1999 | [135] |

决策树以其直观的特征表达、分类准确率高及实现简单等优点，也被研究者应用于网络入侵检测技术中。在误用检测研究中，Kruegel 等人[132]用决策树构建检测模型取代了 Snort [133]中的入侵检测引擎，该检测系统将规则进行聚类从而减少了输入的数据与规则匹配的次数，大大减少了计算时间。最后利用 DARPA 1999 数据集进行实验，结果表明该系统的性能远优于 Snort 的性能。在异常及混合检测研究中，Selim 等人[134]等人提出了采用柔性神经网络与决策树算法相结合多级检测方法对网络恶意行为进行分类，第一级检测区分当前网络行为是恶意行为还是正常行为，如果是恶意行为则进行第二级检测，具体分类为哪一类恶意行为，该方法利用 KDD 1999 数据集作为实验数据取得了平均 93.2%的识别正确率。

此外，无监督学习聚类算法也被用于异常及混合检测中。例如，Blowers 等人[135]使用基于密度的聚类算法 DBSCAN 对正常网络数据包与异常网络数据包进行分组。

利用已知攻击为特征的误用检测，计算量较小、检测精度高，但是无法检测新的攻击类型。面对日益增加的攻击类型，难以及时更新已知攻击特征库，因此难以抵御零日漏洞。基于机器学习的异常及混合检测方法能够发现新的攻击类型，但误判率较高；同时训练需要大量样本且在复杂异构大数据环境下训练难度大，该类方法计算复杂度较高，较难实现实时检测。此外，无论是误用检测还是异常检测在实际网络环境中都较难进行部署实施。究其原因，网络入侵检测与其他应用领域相比具有自身的一些独特性。第一，由于在真实网络环境中，采集网络入侵流量数据比较困难，并且采集耗时较长，因此现有的研究大多采用公开数据集（如表 6 所示）构建网络入侵检测模型，已有研究成果表明，这类模型主要使用网络入侵检测数据验证机器学习算法的性能，而不是基于网络入侵检测技术的实际需求。第二，面对海量的互联网流量数据，很难进行有效的标注。此外，入侵检测系统要求模型要经常再训练以适应日益增加的入侵类型，传统的机器学习技术很难达到对训练模型的自学习的要求。因此，将机器学习技术应用于网络入侵检测，在未来的研究中应着重思考如何安全的获取大量的、有效的网络入侵流量、如何在海量流量中自动提取入侵检测特征，如何构建具有自适应的检测模型三个方面问题。

#### 4.2.3 恶意加密流量识别

基于深度包检测或者模式匹配等方法都对加密流量束手无策，因此识别加密网络流量中包含的威胁是一项具有挑战的工作。2016 年 AISec 会议上 Anderson 等人首次提出在不解密网络流量的情况下，利用机器学习技术从加密的网络流量中识别出具有恶意行为的网络流量[23]，如图 7 所示，首先采集了百万级的正常流量和恶意流量，然后分析了 TLS 流、DNS 流和 HTTP 流的不同之处，具体包括未加密 TLS 握手信息、TLS 流中与目的

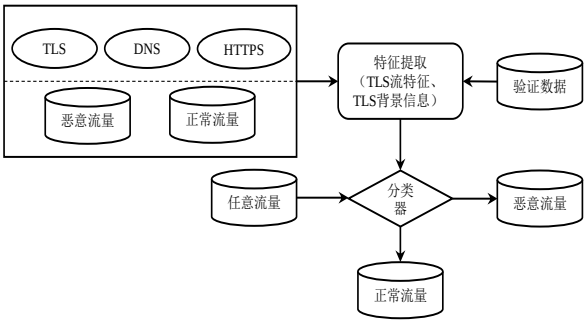


图 7 基于机器学习的恶意加密流量识别流程

IP 地址相关的 DNS 响应信息、相同源 IP 地址 5min 窗口内的 HTTP 流的头部信息；然后从上述信息中提取特征，将该特征采用零均值和单位方差进行归一化处理，随后利用 L1 逻辑回归分类器获得检测模型最优权值，并采用 10 倍交叉验证进行模型验证。最后在实际检测中，将待测加密流量特征向量输入模型，根据预设定的阈值进行判别是否为恶意流量。实验结果显示，该方案对恶意加密流量的识别准确率达到了 99%。该项研究不依赖于加密有效载荷的检测，而利用了 TLS 流相关背景信息（包括 DNS 响应、HTTP 头部等信息）辅助恶意加密流量检测，但该项研究是基于 WindowsXP 系统的 5 分钟窗口采集的流量数据，因此使用该数据集构建的模型可能存在偏差。Conti 等人[137]提出利用机器学习技术分析网络加密流量，用于识别移动终端用户的行为。在不与移动终端交互的前提下，利用已知 APP 在移动终端生成网络流量，在网络侧截取网络流量，将网络流的时间序列进行标记，生成有标签的训练集；然后使用层次聚类（Hierarchical Clustering）算法将网络流进行聚类，相似的流被分组在同一个簇中代表有相似的用户行为，不同的流分组至不同的簇；利用整数形式表示每个簇的特征，再使用随机森林算法执行分类操作，将未知的流量分类至不同的流簇中。尽管该方案针对 Facebook、Gmail、Twitter、Dropbox 等 7 个知名 APP 的数据进行实验均取得了不错的识别效果，但该方案仍存在一定的缺陷：由于训练数据集是人为激发 APP 生成的，与真实网络环境中的数据存在一定的偏差，由此对未在训练数据集中出现的恶恶意加密流量的检测研究近年来刚刚起步，随着恶意流量形式的多样化，未来需要进一步研究加密流量中的典型特征以及其它附加信息；此外还可以尝试

表 7 机器学习在网络安全中的应用

| 网络安全 | 安全问题 | 问题抽象 | 安全特征 | 机器学习算法 | 参考文献 |
|---|---|---|---|---|---|
| 网络基础设施安全 | BGP 异常检测 | 分类/聚类 | BGP 更新消息特征、BGP 的时间序列特征等 | 层次聚类、决策树和 ELM、SVM 和隐马尔可夫模型、朴素贝叶斯、AdaBoost、LSTM | [21,34,98,99,102] |
| | 恶意域名检测 | 分类/聚类 | 基于网络层的特征、基于区域的特征、基于时间的特征、基于 DNS 应答的特征、基于 TTL 的特征、基于域名信息等 | 决策树、X-Means | [46,105,106] |
| 网络安全检测 | 僵尸网络检测 | 分类/聚类 | 快速收敛时间、P2P 节点的通信图、误用检测信息、域名中的语法特征、查询特征、分级域名的统计信息、n 元组的统计信息、加密特征、域名结构化的特征、Bot 意图查询、主机信息、图特征、NetFlow、C&C 通信中的关键特征、C&C 异常命令结合异常日志信息、流量的目的地址和端口聚合集合结合异常日志信息、网络流量数据关联 spam 信息以及 DNS 的日志记录 | X-Means 聚类、图聚类、单链层次聚类算法、SVM、随机森林、隐马尔可夫模型、关联规则 | [109-118] |
| | 网络入侵检测 | 分类/聚类 | 协议标识符（ID）、源端口、目的端口、源地址、目的地址、ICMP 类型、ICMP 代码、原始数据长度、原始数据内容、TCP session 特征、攻击签名、telnet 会话、网络数据包 | 神经网络、关联规则、贝叶斯网络、决策树、隐马尔可夫模型、朴素贝叶斯、SVM、DBSCAN | [12-15,27,50-51,120-122,124-130,132,134-135] |
| | 恶意加密流量识别 | 分类/聚类 | 未加密 TLS 握手信息、DNS 响应信息、HTTP 流的头部信息、网络流的时间序列等 | 逻辑回归、层次聚类、随机森林 | [23,137] |

采用多样化的机器学习模型提升对恶意加密流量识别系统的效率及适应性。

### 4.3 小结

本节主要介绍了机器学习在网络基础设施安全以及网络安全检测中的研究现状，包括 BGP 异常检测、恶意域名检测、僵尸网络的检测、网络入侵检测以及恶意加密流量的识别。这些研究中主要使用的安全特征、机器学习算法以及相关文献如表 7 所示。BGP 路由的安全关乎互联网的连通性以及稳定性，域名系统是各类关键应用的基础，因此网络基础设施安全由于其重要地位对检测的准确率及误报率要求较高。僵尸网络、网络入侵以及恶意加密流量由于攻击流量大、形式多样化，对该类攻击检测要求能够做出快速实时响应。基于机器学习构建在线网络安全检测模型，应满足三方面要求：时间复杂度、增量更新能力及泛化能力。现有的研究中能够达到线性级复杂度的神经网络等算法无法满足对实时流的处理，满足增量更新能力的隐马尔科夫模型及朴素贝叶斯网络等算法却会带来计算的复杂度，具有良好泛化能力的模型则强烈依赖于输入数据集。因此如何选择合适的机器学习算法以及应用于网络安全问题时提高准确率、实时性等均需进一步深入研究。

## 5 机器学习在应用安全研究中的应用

本节从应用软件安全、社会网络安全两个方面，介绍机器学习在应用安全中的相关研究工作。其中，应用软件安全主要包括垃圾邮件检测、PDF 恶意软件检测、恶意网页检测，社会网络安全主要包括社交网络异常帐号检测、信用卡欺诈检测、取证分析、网络舆情。

### 5.1 应用软件安全

电子邮件、PDF、网页等是最常见的软件应用，它们的安全问题已成为学者们关注的热点问题之一。目前，机器学习技术在应用软件安全中的典型研究包括垃圾邮件的检测、恶意网页的识别以及恶意 PDF 文档的检测。

#### 5.1.1 垃圾邮件的检测

传统垃圾邮件检测方法是在服务器端手动设置检测规则，即在服务器端通过修改邮件传输协议、设置发送或接收规则或设置黑白名单等[138]完成垃圾邮件过滤。该方法只能屏蔽已知类型的垃圾邮件，因此检测效率低、规则更新不及时。采用人工干预少且能够自动更新规则的机器学习技术，在服务器端或客户端部署垃圾邮件检测系统，解决了传统垃圾邮件检测方法存在的问题。

垃圾邮件的检测可以抽象为机器学习的文本分类问题，例如，最简单的分类可以定义为{-1,1}，-1 代表非垃圾邮件；1 代表垃圾邮件。为了将垃圾邮件信息能够定义为文本分类问题，首要将垃圾邮件文本信息进行数值化表示，每条消息通常表示为一组向量，特征向量中的元素代表了垃圾邮件中特征值。在这个过程中包含了若干数据预处理的过程，例如垃圾信息的词汇处理、数据清洗、降维、特征表示、归一化等操作，尽量用较少的特征表示垃圾邮件的有效信息。定义特征向量以及构建训练数据集后，选择适当的机器学习分类算法例如朴素贝叶斯分类器、决策树、SVM、神经网络等将待测的邮件与已知垃圾邮件的特征进行匹配，从而识别是否为垃圾邮件。

Stanford 大学的 Sahami[23]最早提出使用朴素贝叶斯分类算法在垃圾邮件检测中。除了使用已有的邮件样本定义的特征之外，还将人工定义的特征添加到特征向量中，例如‘free money’、‘Only $’、域名后缀特征是否有附件以及接收时间等。在统计了不同特征在垃圾邮件与合法邮件中出现的概率之后，利用朴素贝叶斯计算公式取得待测邮件的后验概率。朴素贝叶斯算法遵循样本中的所有特征是相互独立，互不影响的假设条件。在实际情况中，该假设条件难以成立。随后有研究者利用神经网络[16]、SVM[43]等机器学习算法解决基于内容的垃圾邮件检测。

然而垃圾邮件发送者通过更改发送垃圾邮件的 IP 地址或者改变垃圾邮件内容，可以逃避基于内容的垃圾邮件检测系统[139]。因此，有学者提出了不依赖于垃圾邮件内容的检测方法[140]，该方法使用了基于轻量级的网络层的 13 个特征，包含针对单个包的特征，针对单个头部的特征、单个消息的特征以及基于历史信息的聚合特征，利用多个决策树的集成学习方法构建邮件发送者信誉评估系统，将垃圾邮件的发送者 IP 与合法用户区分。该方法仅能处理小规模的邮件数量，扩展性较差。

2008 年之后垃圾邮件过滤技术趋于稳定，鲜见有新的技术出现。随着近年来大数据的不断涌现，如何应对在海量数据中对新型的垃圾邮件快速做出反应仍是亟需思考和解决的问题。此外，垃圾邮

件的识别是典型的在线应用，因此如何能够在邮件识别的过程中自动的实现分类器的更新是未来研究的方向之一。

5.1.2 基于 URL 的恶意网页识别

恶意网页通常指在用户访问网页时能够窃取用户隐私、安装恶意程序或执行恶意代码的网页集合。恶意网页识别通常采用基于黑名单的识别方法、基于规则匹配的方法以及基于主机行为识别的方法。这些方法存在时效性差、误报率高及更新难等问题[141]。机器学习算法以其强大的自学习能力，成为恶意网页识别研究中新的技术路线。目前主要分为基于分类方法的恶意网页识别和基于聚类方法的恶意网页识别。

基于分类方法的恶意网页识别，通常将该问题抽象为机器学习的二分类问题，一般流程如图 8 所示，首先根据已标记 URL 数据集进行特征提取，常用的静态特征包括主机信息、URL 信息和网页信息等，动态特征主要包括浏览器行为、URL 的重定向信息、网页跳转关系等；对上述特征进行归一化处理，归一化后特征的取值用{0，1}表示；已知的网页标记用{0,1}表示，{0}代表正常网页，{1}代表恶意网页；然后选择决策树、贝叶斯网络、SVM、逻辑回归等分类算法构造分类器，进而识别未知类型 URL 数据集。

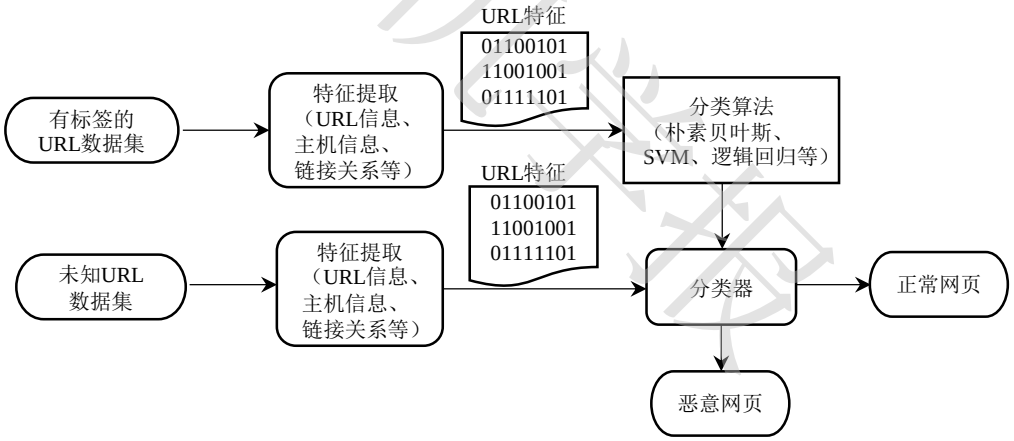


图 8 基于分类方法的恶意网页识别流程

Justin 等人[20]采用 URL 信息的词汇特征以及主机特征，在假定各个特征独立的条件下，利用贝叶斯规则，计算每个特征属于恶意 URL 的概率。对待测的 URL 提取特征后，计算其后验概率，通过预设的阈值来判别待测网页是否属于恶意网页。朴素贝叶斯分类器算法简单，分类速度快，但其局限于要满足于特征独立的假设，然而在恶意网页识别时，该假设条件并不成立。与朴素贝叶斯分类器不同，SVM 算法无需满足特征相互独立的假设，Huang 等人[28]提取了 4 个钓鱼网页 URL 的结构特征、9 个词汇特征及 10 个知名钓鱼网站域名特征，利用 LIBSVM 函数设计钓鱼网页识别系统。虽然该方法对特征直接关系无要求，但 SVM 分类器分类准确性对训练数据和参数依赖较强。SanghoLee 等人[11]从 URL 重定向链接关系出发，引入重定向链接长度、入口 URL 出现频率等信息作为特征，利用逻辑回归分类器学习到一组权值，对可疑 URL 提取特征，将权值与特征向量线性相加后进行归一化处理，进而对可疑 URL 进行识别。该方法分类速度快，但若输入数据有偏差可能导致分类器不收敛。

基于聚类的恶意网页识别的一般流程如图 9 所示，首先将网页采集的 URL 数据集中提取链接关系、URL 特征、网页文本信息等特征，通常采用{0,1}标识；然后根据聚类算法，将 URL 数据集划分为若干聚类，同一聚类的 URL 数据之间具有较高的相似度，而不同聚类的 URL 数据对象之间的相似度较低；最后根据已标记数据的聚类结果，对待测 URL 识别是否是恶意网页。文献[22]从待测网页及关联页面中提取链接、关键词排序、网页文本及层次相似度等关系作为特征，利用已知钓鱼网站 URL 地址集 PhishTank 和基于密度的聚类算法 DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [142]对待测 URL 与已知的钓鱼网页特征进行密度计算，若与已知类别的钓鱼网页接近，即可判别为该类钓鱼网页。该方法特征的提取需要依赖网页文本信息和搜索排序，因此其分类速率较低。

上述经典机器学习算法对恶意网页进行特征提取时，存在分类器速度慢、过拟合或不收敛等问

题；同时海量网页带来的海量特征，容易产生高维特征空间，造成维数灾难；因此需要进一步研究快速有效的特征选择方法以及降维方法对高维数据空间进行处理。此外，在海量的网页数据中，恶意网页与正常网页存在巨大的数据不平衡性[143]，因此如何针对巨大的不平衡数据集对恶意网页进行识别是一项新的挑战。

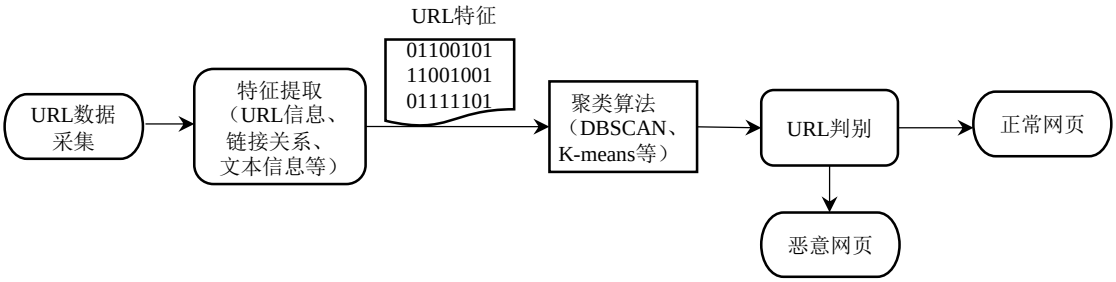


图 9　基于聚类方法的恶意网页识别流程

5.1.3 恶意 PDF 的检测

恶意 PDF 是指在正常的 PDF 文件中嵌入恶意代码。传统的恶意 PDF 检测方法有基于病毒检测[144]、基于签名的的检测方法[145]等，这些方法存在识别率不高、无法及时更新恶意代码等问题。机器学习技术为恶意 PDF 检测提供了新方向，与恶意网页识别类似，恶意 PDF 软件的检测也属于机器学习的二分类问题，但二者采用的特征明显不同。

PDF 文档的检测研究大多采用 PDF 文档内容或结构为特征，利用随机森林、SVM、决策树等分类器构建 PDF 检测器。例如，Charles 等人提出通过随机森林检测含恶意代码的 PDF 文件技术 PDFrate[47]，作者从 PDF 文档元数据以及文档结构中提取了 135 个特征，使用已标记特征的训练数据，并采用 10 倍的交叉验证，生成具有多个分类树的分类器，从待测 PDF 文档提取特征，评估森林中的每个树，最后投票决定其分类。该方法初始训练过程计算开销较大，但一旦分类器构建完成，对待测的 PDF 文档的分类速度很高。PDFrate 仅从简单的 PDF 元数据以及字节级的文件结构提取特征，随后 Nedim 等人提取 PDF 文档对象级的结构特征，构建基于 SVM[146]以及基于决策树的分类器 Hidost[147]，实验表明，该类方法取得较好的分类效果。

然而，上述方法均存在一定的漏洞，攻击者可通过改变 PDF 结构或随机修改恶意代码躲避检测。例如 2016 年 NDSS 会议上，Xu 等[148]采用遗传编程（Genetic Programming）随机修改已知的恶意软件代码，成功躲避了恶意 PDF 文件分类器 PDFrate 和 Hidost 的检测。再如 Nedim 等人提出[149]利用 SVM 生成恶意 PDF 文件，通过改变了 PDF 结构成功躲避了分类器 PDFrate。

从恶意 PDF 文件的攻防博弈两方研究来看，无论以文件内容还是文件结构为特征都不能训练出可靠的 PDF 分类器。究其原因，以文件结构或内容上的差异为特征，训练出来的分类器存在很大盲区。因此，采用机器学习方法构建 PDF 恶意软件检测，在未来的研究中不仅要考究准确率、误报率等度量指标，还应考量更合适的特征，或者利用深度学习等自动提取恶意 PDF 的特征。

### 5.2 社会网络安全

目前，与机器学习相关的社会网络安全研究主要集中在社交网络异常帐号检测、信用卡欺诈检测、取证分析以及网络舆情方面。下面分别详细介绍相关研究工作。

5.2.1 社交网络异常帐号检测

社交网络中存在着大量的虚假帐号和被盗用帐号，这类帐号统称为异常帐号。异常帐号常被用来发布虚假广告、色情、钓鱼等恶意信息[150]，对社交网络用户带来极大的危害。为了高效地挖掘出海量社交网络帐号信息中的异常帐号，已有大量研究利用机器学习技术检测社交网络异常帐号，根据检测特征不同分为基于帐号行为的检测方法、基于消息内容的检测方法。

基于帐号行为的检测方法的关键在于如何选取帐号行为特征及检测算法。基于帐号行为的检测方法的基本流程是，首先在社交网络中获取数据集，构建训练数据集、验证数据集及测试数据集；然后利用异常帐号与正常帐号相比，在发送消息的频率、添加好友的请求等行为方面的差异性，从数据集中提取相应的特征，目前常选取的行为特征有：用户的个人信息、用户行为、帐号创建时间、每天发布消息数量以及好友关系等；然后选择随机

森林、SVM、朴素贝叶斯、K-Means 等算法训练构建分类器。

目前，许多学者从知名社交网络中选取不同的行为特征和机器学习算法进行研究。例如，Stringhini 等人[44]通过网络爬虫收集个人信息作为数据集，从中提取好友请求比率、发布 URL 在日志记录中的比率、消息相似度、好友选择、消息发送数量以及好友数量为特征，使用训练数据训练开源机器学习软件 Weka 中提供的随机森林分类器，采用 10 倍交叉验证方式验证分类器，然后分别对 Facebook、Twitter 中的异常帐号进行检测。随机森林分类器识别准确率高、误报率小，但对于海量的社交网络帐号，随机森林分类器的构建会占用内存大，评估时间较长；同时如果训练数据存在较大噪音时，分类器易过拟合。Wang 等人[151]选取用户点击行为中的 8 个特征和会话级信息的 4 个特征，利用 SVM 构建分类器，对人人网和 LinkedIn 的数据进行正常点击和异常点击行为分类。该方法对未知异常帐号具有较好的检测效果，但由于选取的是帐号长时间点击行为的特征，因此对点击时间较短的帐号存在较高的漏检率。Freeman[152]选取注册时间、点击历史等行为特征，利用朴素贝叶斯分类器对 LinkedIn 网站的异常帐号进行实验，该方法操作简单，但误报率较高。Viswanath 等人[153]针对社交网络用户行为数据的高维度、持续时间长以及存在噪音等问题，将社交网络正常帐号中基于时间、空间和时空的行为作为特征，利用主成分分析 PCA 算法建模，然后将待测帐号的特征后输入该模型，根据偏离程度判断该帐号是否有异常行为。

基于消息内容的异常帐号检测，是根据异常帐号所发布的内容与正常帐号所发布内容的相似程度检测。该方法采用的特征是消息内容本身，例如消息文本中的 URL、消息内容等。目前常采用机器学习有 SVM、逻辑回归分类器、聚类算法等。例如，Egele 等人[154]针对社交网络中帐号被劫持的问题，对采集的数据集中的每条消息选取帐号每日活动小时数、消息来源、消息主题、直接用户交互等 7 个信息为特征，采用开源机器学习软件 Weka 中的 SMO（Sequential Minimal Optimization）算法构建了判别模型，然后对可疑帐号的消息内容进行聚类，根据聚类中已知消息的类型判断帐号是否被劫持。该方法针对消息的相似性进行判别，易于被攻击者躲避检测。Thomas 等人[155]从社交网络帐号提交的每一个 URL 提取特征，包括 URL 的特征、URL 跳转特征、HTML 头部、HTML 内容等，构建逻辑回归分类器来判断是否为恶意的 URL。该方法仅通过 URL 的特征进行判别，因此对无法对其他类型的恶意行为检测。Amleshwaram 等人[156]为了检测 Twitter 中的帐号的异常行为，选取了垃圾邮件的目的域名、推文的内容以及推文的来源为特征，采用 K-Means 算法对 Twitter 中的异常帐号的恶意行为进行检测，该方法仅对具有恶意行为的帐号进行检测，对劫持帐号以及虚假帐号无法检测。此外，还有将帐号行为特征和消息内容结合的异常帐号检测研究。例如 Miller 等人[30]从 Twitter 文本内容、好友及粉丝的帐号信息、好友关系中提取了 107 个特征，利用基于密度的聚类算法 DBSCAN 和 K-Means 聚类算法相结合进行数据流聚类，将正常帐号聚为一类，其他为异常帐号类；但由于实验采用的训练集和测试集数据量较少，并且检测主要依赖于 Twitter 文本内容，因此只能检测发送恶意消息的异常帐号。

**表 8 机器学习在社交网络异常帐号检测中的应用**

| 检测方法 | 主要特征 | 机器学习算法 | 社交网络 | 相关文献 |
|---|---|---|---|---|
| 基于帐号行为的检测方法 | 好友关系、消息内容、URL | 随机森林 | Facebook、Twitter | [44] |
| | 用户点击行为 | SVM | 人人网、LinkedIn | [151] |
| | 用户时间、空间行为 | PCA | Facebook | [152] |
| | 注册时间、点击历史等 | 朴素贝叶斯 | LinkedIn | [153] |
| 基于消息内容的检测方法 | 用户个人信息特征以及微博文本内容 | DBSCAN、K-Means | Twitter | [30] |
| | 时间、消息来源、消息主题、直接用户交互等 | SVM | Twitter | [154] |
| | URL 的特征、URL 跳转的特征、HTML 内容、HTTP 头部、JavaScript 事件、DNS | 逻辑回归 | Twitter | [155] |
| | 推文内容、推文来源、垃圾邮件目的域名 | K-Means | Twitter | [156] |

综上所述，如表 8 所示，社交网络异常帐号检测主要采用随机森林、SVM、朴素贝叶斯及 K-means

等经典的机器学习算法，目前已有的研究主要针对Twitter、LinkedIn、Facebook等知名社交软件数据，根据检测帐号类型不同，提取的特征差异较大。基于帐号行为的异常帐号检测是基于异常帐号的行为特征，因此无法对恶意行为进行实时检测；而基于消息内容的异常帐号检测，在异常帐号发布恶意消息的时候能够及时的检测到，但只是能够检测发布异常消息的帐号，对其他不发布恶意消息的异常帐号无法检测。因此上述两种检测方法均存在着检测模型适应性差，易被攻击者绕过等问题。未来可以利用深度学习、深度增强学习等技术深度挖掘社交网络帐号的行为及内容特征，提升检测模型的自我学习能力。

5.2.2 信用卡欺诈检测

早期，Chan等人[25]在SIGKDD上发表了信用卡欺诈检测研究工作，发现信用卡交易数据巨大、欺诈交易比合法交易少很多、经济损失依赖于交易次数和其他因素等，上述发现体现了信用卡欺诈数据具有稀疏性、非平衡性以及环境复杂性等特性。

Raj等人[157]分析了神经网络、决策树、自组织映射、认知计算、元学习、模式匹配等的信用卡诈骗检测方法，并对比研究其应用效果。Bhattacharyya等人[158]基于国际信用卡组织发布的真实交易数据，并针对该数据的高维性和非平衡性问题，重新生成数据的属性特征值，然后利用SVM、随机森林分别与逻辑回归结合的算法检测信用卡欺诈。此外，信用卡欺诈检测需要较高精度，一旦检测有误会造成不同程度的损失，为此，Sahin等人[159]提出基于代价敏感的决策树方法来检测信用卡欺诈，通过过采样或下采样训练数据来得到代价敏感的属性特征值。Srivastava等人[160]对信用卡交易操作序列建模，使用隐马尔科夫模型对持卡人正常交易行为训练，构建欺诈行为分类器，输入的信用卡交易若不被分类器以较大的概率接受，则被认为是欺诈交易行为。

从现有研究可以看出，信用卡欺诈检测从数据预处理、属性特征选取、机器学习算法选取等方面作了优化研究。然而直接应用机器学习的各类算法并不能取得较好的检测效果，因为信用卡交易数据具有稀疏性和非平衡性等问题，这也是造成现实应用中信用卡检测效果仍不佳的原因。因此训练数据的选取和预处理是信用卡欺诈检测研究的难点，未来可利用聚类等方法识别隐藏在数据中的特征属性，并进一步过滤掉无用数据。此外，面对不断增长的金融大数据，交易欺诈数据的稀疏性、非平衡性以及金融环境的复杂性等问题将变得更为复杂，传统的人工特征和机器学习方法进行信用卡欺诈检测将面临更大的挑战。未来研究可尝试利用先进的机器学习技术提高信用卡欺诈检测系统的准确率，例如利用深度学习自动提取信用卡交易特征属性值；以及针对实际应用场景，还可从进一步地提高信用卡欺诈检测系统的实时性方面着手研究，并在检测准确性和实时性上实现突破。

5.2.3 取证分析

取证分析常被用来调查诈骗、侵占知识产权、入侵攻击等犯罪行为。常见的取证技术包括电子取证、计算机取证和网络取证，相应的证据包括电子设备、计算机、网络运行过程中反映事实的数字信息或数据。典型的数字证据有存储设备中的图像或音视频文件、计算机系统的日志、网络入侵检测系统的工作记录等。传统人工分析数字证据的方式存在时间长、过程复杂等的问题，且需要分析的证据数据量越来越大，因此已有部分研究开始利用机器学习进行取证分析。

Pearl等人[161]利用机器学习技术识别文件中的伪造笔迹，主要将书写特征、笔迹内容特征为输入，采用监督学习中的稀疏多项式逻辑回归分类器进行笔迹识别。Khanna等人[162]将相机、扫描仪或计算机图形软件生成的图像的差值作为残差模式噪声特征，利用SVM分类器区分是否为合成图像。Khan等人[163]利用贝叶斯网络从文件系统活动（例如访问、创建、修改、删除）中分析证据。Mukkamala等人[164]提出利用ANN和SVM来识别大量信息中的有效证据。Palomo等人利用网络流量日志进行数字证据挖掘，采用SOM算法仅挖掘出流量数据的定性特征[165]，随后基于增长式层次自组织映射网络（Growing Hierarchical Self-organizing Maps，GHSOM）分析高维的日志数据[166]，该方法不仅能提取定性特征，还能提取定量特征。

目前基于机器学习的取证分析处于起步阶段，大多数研究仅提供了一种技术方案，存在无法解释事件因果关系及取证过程复杂、分析时间长等问题。因此，未来取证分析技术可以基于深度学习、迁移学习等进一步提升数字取证效果；还可以研究追踪溯源取证系统，以提升基于机器学习的取证技术的可信度。

5.2.4 网络舆情

网络舆情是在网络空间下网民对事件的态度、

意见及其影响力的集合[167]。国外最早的相关研究为 1996 年美国国防部等提出的 TDT（Topic Detection and Tracking）项目[168]，主要目的是探索新话题出现和追踪再现它们的演变。大量研究诸如热点识别[167]、追踪及趋势分析[169]、观点挖掘与情感分析[170-171]均集中在数据挖掘和信息检索领域，本文从安全角度出发介绍具有代表性的网络舆情研究工作。

在危险事件识别方面，Alsaedi 等人[172]提出公共危害事件实时识别框架，整个流程如图 10 所示，简单概括为：收集 Twitter 的用户生成内容数据；数据预处理；利用朴素贝叶斯分类模型来区分是否是危害事件；选取时间、空间、文本等特征，再利用在线聚类算法得出公共的危害事件；最后生成危害事件摘要信息。该研究不需要危害事件的先验知识，且与 SVM、逻辑回归等分类算法相比，文中选取的朴素贝叶斯算法效果最佳。在舆情分析方面，Liu 等人[173] 针对舆情数据的半结构化特性，

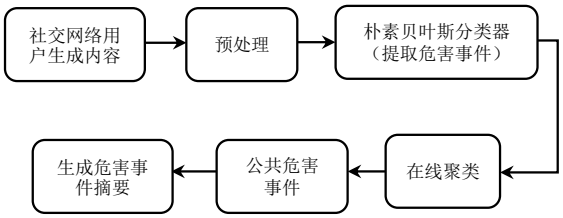


图 10 危险事件识别流程图

提出用传统的向量模型来表示网络舆情的文本格式，利用 K 均值算法对从网站收集的语料库实现文本聚类，并使用 SVM 分类器对新发现的舆情文本分类。针对热点检测，Li 等人[174]提出基于文本挖掘和情感分析的在线论坛热点检测与预测，将 K 均值算法和 SVM 分别作为自动分析文本情感的方法。实验数据取自新浪体育论坛，结果显示 SVM 与 K 均值算法预测结果类似。针对网络舆情预测，Zeng 等人 [175]将隐马尔科夫模型应用到舆情预测，构建了描述网络舆情状态和特征的数学模型，该模型能够根据预测错误率动态调整模型参数。

表 9 机器学习在应用安全中的应用

| 应用安全 | 安全问题 | 问题抽象 | 安全特征 | 机器学习算法 | 参考文献 |
|---|---|---|---|---|---|
| 应用软件安全 | 垃圾邮件检测 | 分类/降维 | 邮件的发送者 IP 地址、邮件内容文本特征、域名特征 | 朴素贝叶斯、神经网络、SVM、决策树集成法 | [16,24,43] |
| | 基于 URL 的恶意网页识别 | 分类/聚类/降维 | 主机信息、URL 信息、网页信息、浏览器行为、URL 的重定向信息、网页跳转关系 | 决策树、贝叶斯网络、SVM、逻辑回归、DBSCAN | [11,20,22.28] |
| | 恶意 PDF 检测 | 分类 | PDF 文档内容、PDF 文件结构 | 随机森林、SVM、决策树、遗传编程 | [47,146-149] |
| 社会网络安全 | 社交网络帐号异常检测 | 分类/聚类/降维 | 用户的个人信息、用户行为、帐号创建时间、每天发布消息数量以及好友关系、消息文本中的 URL、消息内容等 | 随机森林、SVM、PCA、朴素贝叶斯、逻辑回归、K-Means、DBSCAN | [30,44,151-156] |
| | 信用卡欺诈检测 | 分类 | 信用卡交易次数、行为等 | 神经网络、决策树、SVM、随机森林、隐马尔科夫模型 | [157-160] |
| | 取证分析 | 分类/降维 | 书写特征、笔迹内容特征、残差模式噪声特征、文件系统活动、网络流量日志 | 逻辑回归、ANN、SVM、SOM | [161-166] |
| | 网络舆情 | 分类/聚类 | 用户生成内容的时间、空间、文本等特征 | 朴素贝叶斯、K 均值、SVM、隐马尔科夫模型 | [172-175] |

现有的基于机器学习的网络舆情研究，主要针对了具有半结构化数据特征的文本，而视频、语音等非结构化数据的网络舆情分析、预测等研究较少，因此进一步加强视频、语音舆情的研究将会使整个网络舆情分析更为完善。在网络舆情安全势态的实时感知、实时预测方面，现有的技术方案的效率仍较低，未来还需进一步研究如何提高效率。总体而言，专门针对安全领域的网络舆情研究还较少，随着国内外网络空间安全战略的发布，网络舆情将受到更多的关注，特别是公共安全的网络舆情

发现、预测、响应的研究，因此，相关的网络舆情技术将是未来热门研究方向之一。

### 5.3 小结

本节主要介绍了机器学习在应用安全领域的研究现状，包括应用软件安全、社会网络安全两个方面，其中应用软件安全和社会网络安全常用的安全特征、机器学习算法以及相关文献如表 9 所示。

在应用软件安全研究中，主要利用机器学习技术对应用软件的信息内容作出安全防护，从邮件的文本内容、网页信息或 PDF 文件内容、结构，构建能够自动更新的检测模型，但现有检测模型过度依赖于训练数据，泛化能力较差，因此需要深入挖掘应用软件的安全特征，提高检测模型的泛化能力。在社会网络安全方面，针对社交网络中异常账号检测，需要进一步深入挖掘账号及内容特征，构建更具泛化能力的模型。针对信用卡欺诈检测，需要进一步研究信用卡欺诈行为检测中的数据集稀疏性、非平衡性以及金融环境的复杂性等问题。针对取证分析，需要解决因果关系不可解释的问题。针对网络舆情，需要提高网络舆情安全势态感知、预测的实时性。

## 6 研究展望与挑战

通过上述分析可知，目前基于机器学习的网络空间安全研究在系统安全、网络安全及应用安全领域中已有不少解决方案和方法，在包括硬件木马检测、网络入侵检测、社交网络帐号检测等领域均取得了不错的检测效果。但是无论是模型的泛化能力，还是检测准确度、实时性等问题，目前的技术解决方案均不能完全满足网络空间安全的应用需求，并存在一些目前难以解决的问题以及可进一步研究的方向，采用机器学习技术解决网络空间安全问题仍是极具挑战性的工作。同时机器学习技术本身存在一定的研究难点，在解决网络空间安全问题中面临巨大挑战：

（1）基于机器学习的安全解决方案的可解释性与溯源性

大多数机器学习的算法都是黑盒模型，其学习算法可能是公开和透明的，但它产生的模型却是不可解释的[176-178]。例如对恶意代码分类的结论无法获知其原由，尤其当算法出现错判时，很难判别是因为模型参数不正确还是神经网络中某个神经元出现了问题。此外，在网络空间安全问题中，常常要求对安全的源头进行追溯，机器学习技术的解决方案具有无法溯源的根本问题。因此，在未来的研究中，如何能够使机器学习的解决方案具备可解释性、鲁棒性，同时如何能够对网络空间安全问题进行溯源是一个亟待解决的问题。

（2）基于机器学习技术的攻击的防御难度

目前机器学习技术被广泛的应用在系统硬件检测、用户身份认证、恶意域名检测、垃圾邮件检测、恶意软件检测等安全检测研究领域。与此同时，攻击者也开始研究如何利用机器学习技术进行有效的网络攻击[19,65,84,140,148]。例如文献[179]中提出了一种基于梯度的方法构造攻击性的数据，将少量攻击性数据放入训练样本中，用于 SVM 的训练，会导致 SVM 在检测样本上的错误率明显提升。使用机器学习构建的攻击工具会轻易的躲避现有的基于机器学习的防御设施。采用机器学习技术作为攻击技术，增加了检测及防御的难度，目前的研究中对该类攻击的解决方案较少。在未来的研究中，针对网络空间安全数据的不断增加以及攻击种类多样化的现状，如何能够提高实时监测效率以及解决基于机器学习技术的攻击有待研究者进一步深入研究。

（3）机器学习自身的安全问题

机器学习构建的模型自身也并非是一定安全的[180-182]。例如，大量研究采用机器学习构建训练模型，但其输入数据中包含大量的隐私数据，如果攻击者对该模型进行分析，极易得到用户的敏感数据。此外，最近的一些研究中对机器学习构建的模型的输入样本进行攻击，从而使机器学习模型无法取得良好的检测效果。例如增强学习算法 DQN, TRPO 以及 A3C 等都能够轻易的被对抗样本所操控[183-184]，将对抗样本作为输入，即使其中仅包含人类难以察觉的轻微的扰动，也会导致极大的系统性能下降。对抗样本作为机器学习模型的输入，使得系统难以防御它们。2014 年 Goodfellow 提出的生成对抗网络[41]，将生成模型与识别模型相结合，采用博弈思想训练识别模型，为提高系统安全性能提供了一种可能性，但是生成对抗网络如何与网络空间安全领域问题结合仍需进一步的研究。

## 7 结束语

网络空间安全研究网络空间中的安全威胁和防护问题，不仅关系到国家安全，更与人们的日常

生活息息相关。网络空间中存在着大量的网络流量、日志信息、系统信号等数据，深入挖掘这些数据的特征及关联关系，能够为网络空间各级应用提供安全防护措施。机器学习作为当前人工智能领域最热门的研究方向之一，在图像识别、语音识别等领域取得了一系列的令人瞩目的研究成果，吸引了越来越多网络空间安全领域研究人员的关注，取得了一系列的重要研究成果。本文对这些成果进行了系统的总结和分析，以机器学习为技术手段，着重总结了机器学习如何应用于网络空间安全领域，对机器学习在网络空间中的系统安全、网络安全以及应用安全问题进行了深入分析和比较，最后探讨了机器学习技术在网络空间安全领域的发展趋势及挑战。

**致 谢** 衷心感谢评审专家和计算机学报的编辑们对本文提出的宝贵意见和建议！ 致谢

## 参 考 文 献

[1] Luo J, Yang M, Ling Z, et al. Architecture and key technologies of cyberspace security. SCIENTIA SINICA Informationis, 2016, 46(8): 939-968 (in Chinese)
(罗军舟, 杨明, 凌振等. 网络空间安全体系与关键技术. 中国科学: 信息科学, 2016, 46(8): 939)

[2] Li J, Qiu W, Meng K, and Wu Jun. Discipline construction and talents training of cyberspace security. Journal of Information Security Research, 2015, 1(2): 149-154 (in Chinese)
(李建华, 邱卫东, 孟魁等. 网络空间安全一级学科内涵建设和人才培养思考. 信息安全研究, 2015, 1(2): 149-154)

[3] Denning D E. An Intrusion-Detection Model. IEEE Press, 1987.

[4] Jiang H, Nagra J, Ahammad P. SoK: Applying machine learning in security-a survey. arXiv preprint arXiv:1611.03186, 2016

[5] Buczak A L, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 2016, 18(2): 1153-1176

[6] Sommer R, Paxson V. Outside the closed world: on using machine learning for network intrusion detection//Proceedings of the 2010 IEEE Symposium on Security and Privacy. Washington, USA, 2010: 305-316

[7] Nishani L, Biba M. Machine learning for intrusion detection in MANET: a state-of-the-art survey. Journal of Intelligent Information Systems, 2016, 46(2): 391-407

[8] Zhou Z. Machine learning. Tsinghua University Press, 2016(in Chinese)
(周志华. 机器学习. 北京: 清华大学出版社, 2016)

[9] Li Z, Wang W, Wilson C, et al. Fbs-radar: Uncovering fake base stations at scale in the wild//Proceedings of the Network and Distributed System Security Symposium 2017. San Diego, USA, 2017, 1-15

[10] Wang Z. The applications of deep learning on traffic identification. Report: BlackHat, USA, 2015

[11] Lee S, Kim J. WarningBird: A near real-time detection system for suspicious URLs in Twitter stream. IEEE Transactions on Dependable and Secure Computing, 2013, 10(3): 183-195

[12] Bivens A, Palagiri C, Smith R, Szymanski B, Embrechts M. Network-based intrusion detection using neural networks. Intelligent Engineering Systems through Artificial Neural Networks, 2002, 12(1): 579–584

[13] Lippmann, Richard P, Cunningham, et al. Improving intrusion detection performance using keyword selection and neural networks. Computer Networks, 2000, 34(4): 597-603

[14] Brahmi H, Brahmi I, Yahia S B. OMC-IDS: At the cross-roads of OLAP mining and intrusion detection//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin Heidelberg, 2012: 13-24

[15] Li Y, Xia J, Zhang S, et al. An efficient intrusion detection system based on support vector machines and gradually features removal method. Expert Systems with Applications, 2012, 39(1): 424-430

[16] Wu C H. Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications, 2009, 36(3): 4321-4330

[17] Seifert C, Welch I, Komisarczuk P, et al. Identification of malicious web pages through analysis of underlying DNS and web server relationships//Proceedings of IEEE Conference of Local Computer Networks. Montreal, Canada, 2008: 935-941

[18] Tzeng M F. Routing table partitioning for speedy packet lookups in scalable routers. IEEE Transactions on Parallel & Distributed Systems, 2006, 17(5): 481-494

[19] Filar B, Filar B, Filar B. DeepDGA: adversarially-tuned domain generation and detection//Proceedings of ACM Workshop on Security and Artificial Intelligence. New York, USA, 2016:13-21

[20] Ma J, Saul L K, Savage S, et al. Voelker. Beyond blacklists: learning to detect malicious web sites from suspicious URLs//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2009: 1245-1253

[21] Li Y, Xing H J, Hua Q, Wang X Z, Batta P, Haeri S, Trajkovic L. Classification of BGP anomalies using decision trees and fuzzy rough sets//Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. San Diego, USA, 2014: 1331-1336

[22] Liu G, Qiu B, Wenyin L. Automatic detection of phishing target from phishing webpage//Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 4153-4156

[23] Anderson B, Mcgrew D. Identifying encrypted malware traffic with contextual flow data//Proceedings of ACM Workshop on Security and Artificial Intelligence. New York, USA, 2016: 35-46

[24] Sahami M, Dumais S, Heckerman D, et al. A bayesian approach to filtering junk e-mail//Proceedings of AAAI Workshop on Learning for Text Categorization. Madison, USA, 1998: 55-62

[25] Chan P K, Stolfo S J. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection//Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York, USA, 1998: 164-168

[26] Arp D, Spreitzenbarth M, Hubner M, et al. DREBIN: effective and explainable detection of android malware in your pocket//Proceedings of the Network and Distributed System Security Symposium 2014. San Diego, USA, 2014: 1-15

[27] Cannady J. Artificial neural networks for misuse detection//Proceedings of the 1998 National Information Systems Security Conference. Arlington, USA, 1998: 443–456

[28] Huang H, Qian L, Wang Y. A SVM-based technique to detect phishing URLs. Information Technology Journal, 2012, 11(7): 921-925

[29] Hinton, G. E, Osindero, S, Teh, Y. A fast learning algorithm for deep belief nets. Neural Computation, 2006, 18: 1527-1554

[30] Miler Z, Dickinson B, Deitrick W, et al. Twitter spammer detection using data stream clustering. Information Sciences, 2014, 260: 64-73

[31] Jung W, Kim S, Choi S. Poster: Deep learning for zero-day flash malware detection//Proceedings of IEEE Symposium on Security and Privacy. San Jose, California, 2015

[32] Yuan Z, Lu Y, Wang Z, et al. Droid-Sec: deep learning in android malware detection. ACM Sigcomm Computer Communication Review, 2014, 44(4): 371-372

[33] Javaid A, Niyaz Q, Sun W, et al. A deep learning approach for network intrusion detection system//Proceedings of EAI International Conference on Bio-Inspired Information and Communications Technologies. New York, USA, 2015: 21-26

[34] Cheng M, Xu Q, Lv J, et al. MS-LSTM: A multi-scale LSTM model for BGP anomaly detection//Proceedings of International Conference on Network Protocols. Singapore, 2016: 1-6

[35] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy//Proceedings of the 23rd ACM Conference on Computer and Communications Security. Vienna, Austria, 2016: 308-318

[36] Shokri R, Shmatikov V. Privacy-preserving deep learning//Proceedings of the 22nd ACM Conference on Computer and Communications Security. Denver, USA, 2015: 1310-1321

[37] Pan S J, Yang Q. A Survey on transfer learning. IEEE Educational Activities Department, 2010, 22(10): 1345-1359

[38] Liu Y, Huang K, Makris Y. Hardware Trojan detection through golden chip-free statistical side-channel fingerprinting//Proceedings of the 51st Annual Design Automation Conference. San Francisco, USA, 2014: 1-6

[39] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J., Bellemare M. G., et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

[40] Li Y, Liu J, Li Q, Xiao L. Mobile cloud offloading for malware detections with learning//Proceedings of the 34th Annual IEEE International Conference on Computer Communications. Hong Kong, China, 2015: 197–201

[41] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of International Conference on Neural Information Processing Systems. Montréal, Canada, 2014: 2672-2680

[42] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN. arXiv:1702.05983, 2017

[43] Haider P, Brefeld U, Scheffer T. Supervised clustering of streaming data for email batch detection//Proceedings of International Conference on Machine Learning. Corvallis, USA, 2007: 345-352

[44] Stringhini G, Kruegel C, Vigna G. Detecting spammers on social networks//Proceedings of ACM Computer Security Applications Conference. Austin, USA, 2010: 1-9

[45] Danev B, Heydt-Benjamin T S, Capkun S. Physical-layer identification of RFID devices//Proceedings of the 18th Conference on USENIX Security Symposium. Montreal, 2009: 199–214

[46] Antonakakis M, Perdisci R, Dagon D, et al. Building a dynamic reputation system for DNS//Proceedings of the 19th USENIX Security Symposium. Washington, USA, 2010: 273-290

[47] Smutz C, Stavrou A. Malicious PDF detection using metadata and structural features//Proceedings of the 28th Annual Computer Security Applications Conference. Orlando, USA, 2012: 239-248

[48] Jap D, He W, Bhasin S. Supervised and unsupervised machine learning for side-channel based Trojan detection//Proceedings of the IEEE 27th International Conference on Application-specific Systems, Architectures and Processors. London, England, 2016: 17-24

[49] Narudin F A, Feizollah A, Anuar N B, et al. Evaluation of machine learning classifiers for mobile malware detection. Soft Computing, 2016, 20(1): 343-357

[50] Tajbakhsh A, Rahmati M, Mirzaei A. Intrusion detection using fuzzy association rules. Applied Soft Computing, 2009, 9(2): 462-469

[51] Amor N B, Benferhat S, Elouedi Z. Naive Bayes vs decision trees in intrusion detection systems//Proceedings of ACM Symposium on Applied Computing. Nicosia, Cyprus, 2004: 420-424

[52] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8):861-874

[53] Gu GF, Zhang JJ, Lee WK. BotSniffer: Detecting botnet command and control channels in network traffic//Proceedings of Annual Network & Distributed System Security Symposium. San Diego, USA, 2008: 1-18

[54] Rostami M, Koushanfar F, Karri R. A primer on hardware security: Models, methods, and metrics. Proceedings of the IEEE, 2014, 102(8): 1283-1295

[55] Guin U, Huang K, DiMase D, et al. Counterfeit integrated circuits: a rising threat in the global semiconductor supply chain. Proceedings of the IEEE, 2014, 102(8): 1207-1228

[56] Huang K, Liu Y, Korolija N, et al. Recycled IC detection based on statistical methods. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015, 34(6): 947-960

[57] Xiao K, Forte D, Tehranipoor M M. Circuit Timing Signature (CTS) for detection of counterfeit integrated circuits. Springer International Publishing, 2016: 211-239

[58] Asadizanjani N, Dunn N, Gattigowda S, et al. A database for counterfeit electronics and automatic defect detection based on image processing and machine learning//Proceedings of the 42nd International Symposium for Testing and Failure Analysis. Texas, USA, 2016: 1-8

[59] Tehranipoor M, Koushanfar F. A survey of hardware trojan taxonomy and detection. IEEE Design & Test of Computers, 2010, 27(1): 10-25

[60] Wang C, Jiang P, Yu M. Survey of hardware Trojan horse detection on chip. Semiconductor Technology, 2012: 341-346 (in Chinese)
(王晨旭, 姜佩贺, 喻明艳. 芯片级木马检测技术研究综述. 半导体技术, 2012: 341-346)

[61] Bao C, Forte D, Srivastava A. On application of one-class SVM to reverse engineering-based hardware Trojan detection//Proceedings of 15th International Symposium on Quality Electronic Design. Santa Clara, USA, 2014: 47-54

[62] Iwase T, Nozaki Y, Yoshikawa M, et al. Detection technique for hardware Trojans using machine learning in frequency domain//Proceedings of 2015 IEEE 4th Global Conference on Consumer Electronics. Osaka City, Japan, 2015: 185-186

[63] Gassend B, Clarke D, Van Dijk M, et al. Silicon physical random functions//Proceedings of the 9th ACM conference on Computer and Communications Security. Washington, USA, 2002: 148-160

[64] Rührmair U, Sehnke F, Sölter J, et al. Modeling attacks on physical unclonable functions//Proceedings of the 17th ACM conference on Computer and Communications Security. Chicago, USA, 2010: 237-249

[65] Hospodar G, Maes R, Verbauwhede I. Machine learning attacks on 65nm Arbiter PUFs: accurate modeling poses strict bounds on usability//Proceedings of 2012 IEEE International Workshop on Information Forensics and Security. Tenerife, Spain, 2012: 37-42

[66] Tekbas O¨H , Serinken N, U¨reten O . An experimental performance evaluation of a novel radio-transmitter identification system under diverse environmental conditions. Canadian Journal of Electrical and Computer Engineering, 2004, 29(3): 203-209

[67] Brik V, Banerjee S, Gruteser M, et al. Wireless device identification with radiometric signatures//Proceedings of the 14th ACM International Conference on Mobile Computing and Networking. San Francisco, USA, 2008:116-127

[68] Dey S, Roy N, Xu W, et al. AccelPrint: imperfections of accelerometers make smartphones trackable//Proceedings of the Network and Distributed System Security Symposium 2014. San Diego, USA, 2014: 1–16

[69] Hospodar G, Gierlichs B, De Mulder E, et al. Machine learning in side-channel analysis: a first study. Journal of Cryptographic Engineering, 2011, 1(4): 293-302

[70] Lerman L, Medeiros S F, Veshchikov N, et al. Semi-supervised template attack//Proceedings of International Workshop on Constructive Side-Channel Analysis and Secure Design. Berlin Heidelberg, 2013: 184-199

[71] Lerman L, Bontempi G, Markowitch O. Power analysis attack: an approach based on machine learning. International Journal of Applied Cryptography, 2014, 3(2): 97-115

[72] van Do Thanh, Nguyen H T, Momchil N. Detecting IMSI-Catcher using soft computing//Proceedings of International Conference on Soft Computing in Data Science. Singapore, 2015: 129-140

[73] Yamaguchi F, Lindner F, Rieck K. Vulnerability extrapolation: assisted discovery of vulnerabilities using machine learning//Proceedings of USENIX Workshop on Offensive Technologies. San Francisco, USA, 2011: 13-13

[74] Shin E C R, Song D, Moazzezi R. Recognizing functions in binaries with neural networks//Proceedings of the 24th USENIX Security Symposium. Washington, USA, 2015: 611-626

[75] Yamaguchi F, Wressnegger C, Gascon H, et al. Chucky: exposing missing checks in source code for vulnerability discovery//Proceedings of the 2013 ACM conference on Computer & Communications Security. Berlin, Germany, 2013: 499-510

[76] Pang Y, Xue X, Namin A S. Early identification of vulnerable software components via ensemble learning//Proceedings of 2016 15th IEEE International Conference on Machine Learning and Applications. Orange, USA, 2016: 476-481

[77] Scandariato R, Walden J, Hovsepyan A, et al. Predicting vulnerable software components via text mining. IEEE Transactions on Software Engineering, 2014, 40(10): 993-1006

[78] Pang Y, Xue X, Namin A S. Predicting Vulnerable software components through N-Gram analysis and statistical feature selection//Proceedings of 2015 14th IEEE International Conference on Machine Learning and Applications. Miami, USA, 2015: 543-548

[79] Long F, Rinard M. Prophet: automatic patch generation via learning from successful patches. Cambridge, Massachusetts: CSAIL, Technical Report: MIT-CSAIL-TR-2015-027, 2015

[80] Nath H V, Mehtre B M. Static malware analysis using machine learning methods//Proceedings of International Conference on Security in Computer Networks and Distributed Systems. Berlin Heidelberg, 2014: 440-450

[81] Nissim N, Moskovitch R, Rokach L, et al. Novel active learning methods for enhanced PC malware detection in windows OS. Expert Systems with Applications, 2014, 41(13): 5843-5857

[82] Wilhelm J, Chiueh T. A forced sampled execution approach to kernel rootkit identification//Proceedings of International Workshop on Recent Advances in Intrusion Detection. Berlin Heidelberg, 2007: 219-235

[83] Chen S, Xue M, Tang Z, et al. Stormdroid: a streaminglized machine learning-based system for detecting android malware//Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. Xi'an, China, 2016: 377-388

[84] Golle P. Machine learning attacks against the Asirra CAPTCHA//Proceedings of the 15th ACM conference on Computer and Communications Security. Alexandria, USA, 2008: 535-542

[85] Yue Q, Ling Z, Fu X, et al. Blind recognition of touched keys on mobile devices//Proceedings of the 2014 ACM Conference on Computer and Communications Security. Scottsdale, USA, 2014: 1403-1414

[86] Liu, Xiangyu, et al. When good becomes evil: keystroke inference with smart watch//Proceedings of the 22nd ACM Conference on Computer and Communications Security. Denver, USA, 2015: 1273-1285

[87] Zheng, Nan, et al. You are how you touch: user verification on smartphones via tapping behaviors//Proceedings of 2014 IEEE 22nd International Conference on Network Protocols. North Carolina. USA, 2014: 221-232

[88] Giuffrida C, Majdanik K, Conti M, et al. I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics//Proceedings of International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Egham. London, UK, 2014. 92-111

[89] Deng Y, Zhong Y. Keystroke dynamics advances for mobile devices using deep neural network. Recent Advances in User Authentication Using Keystroke Dynamics Biometrics, 2015, 2: 59-70

[90] Kobojek P, Saeed K. Application of recurrent neural networks for user verification based on keystroke dynamics. Journal of Telecommunications and Information Technology, 2016, (3): 80

[91] Li L, Zhao X, Xue G. Unobservable re-authentication for smartphones//Proceedings of the Network and Distributed System Security Symposium 2013. San Diego, USA, 2013: 1-16

[92] Green M. The threat in the cloud. IEEE Security & Privacy, 2013, 11(1): 86-89

[93] Zhang Y, Juels A, Reiter M K, et al. Cross-VM side channels and their use to extract private keys//Proceedings of the 2012 ACM conference on Computer and Communications Security. Raleigh, USA, 2012: 305-316

[94] Gulmezoglu B, Eisenbarth T, Sunar B. Cache-based application detection in the cloud using machine learning//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 288-300

[95] Fischer A, Kittel T, Kolosnjaji B, et al. Cloudidea: A malware defense architecture for cloud data centers//Proceedings of OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Cham, 2015: 594-611

[96] Wang Na, Du xuehui, Wang Wenjuan, Liu aodi. A survey of border gatewary protocol security. Chinese Journal of Computers, 2017, 40(7): 1626-1648 (in Chinese)
(王娜，杜学绘，王文娟，刘敖迪．边界网关协议安全研究综述．计算机学报, 2017, 40(7): 1626-1648)

[97] Basseville M, Nikiforov I V. Detection of abrupt changes: theory and application. Englewood Cliffs: Prentice Hall, 1993

[98] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications. Neuro computing, 2006, 70(1): 489-501

[99] Al-Rousan N M, Trajkovic L. Machine learning models for classification of BGP anomalies//Proceedings of IEEE 13th International Conference on High Performance Switching and Routing. Belgrade, Serbia, 2012: 103–108

[100] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques. ACM SIGMETRICS Performance Evaluation Review, 2005, 33(1): 50-60

[101] Hu W, Hu W, Maybank S. Adaboost-based algorithm for network intrusion detection. IEEE Transactions on Systems Man & Cybernetics- Part B Cybernetics, 2008, 38(2): 577-583

[102] Qiu T, Ji L, Pei D, Wang J, Xu J J, Ballani H. Locating prefix hijackers using lock//Proceedings of 18th Conference on USENIX Security Symposium. Montreal, Canada, 2009: 135-150

[103] Villamarin-Salomon R, Brustoloni J C. Identifying botnets using anomaly detection techniques applied to DNS traffic//Proceeding of IEEE Consumer Communications and Networking Conference. Las Vegas, USA, 2009: 476-481

[104] Ricardo Villamarín-Salomón. Bayesian bot detection based on DNS traffic similarity//Proceedings of ACM Symposium on Applied Computing. Hawaii, USA, 2009: 2035-2041

[105] Bilge L, Kirda E, Kruegel C, et al. EXPOSURE: finding malicious domains using passive DNS analysis// Proceedings of Annual Network & Distributed System Security Symposium. San Diego, USA, 2011: 1-17

[106] Bilge L, Sen S, Balzarotti D, et al. Exposure: a passive DNS analysis service to detect and report malicious domains. ACM Transactions on Information & System Security, 2014, 16(4): 1-28

[107] Rafael A Rodriguez-Gomez, Gabriel Macia-Fernandez, Pedro Garcia-Tedoro. Survey and taxonomy of botnet research through life-cycle. ACM Computing Surveys, 2013, 45(4): 1-33

[108] Jiang Jian, Zhuge Jianwei, Duan Haixin, Wu Jianping. Research on botnet mechanisms and defenses. Journal of Software, 2012, 23(1): 82-96 (in Chinese)
(江健，诸葛建伟，段海新，吴建平．僵尸网络机理与防御技术．软件学报, 2012, 23(1): 82-96)

[109] Nagaraja S, Mittal P, Hong CY, et al. BotGrep: finding P2P bots with structured graph analysis//Proceedings of 19th USENIX Conference on Security Symposium. Washington, USA, 2010: 7-7

[110] Antonakakis M, Perdisci R, Nadji Y, et al. From throw-away traffic to bots: detecting the rise of DGA-based malware//Proceedings of 21th

USENIX Conference on Security Symposium. Bellevue, USA, 2012: 1-16

[111] Zhang J, Xie Y, Yu F., Soukal D, Lee W. Intention and origination: an inside look at large-scale bot queries//Proceedings of Annual Network & Distributed System Security Symposium 2013. San Diego, USA, 2013: 1-16

[112] Jacob G, Hund R, Kruegel C, et al. JACKSTRAWS: picking command and control connections from bot traffic//Proceedings of 20th USENIX Conference on Security Symposium. San Diego, USA, 2011: 1-16

[113] Bilge L, Balzarotti D, Robertson W, et al. Disclosure: detecting botnet command and control servers through large-scale netflow analysis//Proceedings of the 28th Annual Computer Security Applications Conference. Orlando, USA, 2012: 129-138

[114] Chen F, Ranjan S, Tan P N. Detecting bots via incremental LS-SVM learning with dynamic feature adaptation//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 386-394

[115] Gu GF, Zhang JJ, Lee WK. BotSniffer: Detecting botnet command and control channels in network traffic//Proceedings of Annual Network & Distributed System Security Symposium. San Diego, USA, 2008: 1-18

[116] Gu GF, Perdisci R, Zhang JJ, Lee WK. BotMiner: clustering analysis of network traffic for protocol and structure-independent botnet detection//Proceedings of 17th USENIX Conference on Security Symposium. San Jose, USA, 2008: 139−154

[117] Hu X, Knysz M, Shin KG. Rb-Seeker: auto-detection of redirection botnets//Proceedings of Annual Network & Distributed System Security Symposium. San Diego, USA, 2009: 1-17

[118] Tegeler F, Fu Xiaoming, Vigna G, et al. BotFinder: finding bots in network traffic without deep packet inspection//Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies. New York, USA, 2012: 349-360

[119] Morel B. Artificial intelligence and the future of cybersecurity//Proceedings of the 4th ACM workshop on Security and artificial intelligence. Chicago, USA, 2011: 93-98

[120] Apiletti D, Baralis E, Cerquitelli T, D'Elia V. Characterizing network traffic by means of the NetMine framework. Computer Networks, 2009, 53(6): 774-789

[121] Luo J, Bridges S. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. International Journal of Intelligent Systems, 2000, 15(8): 687–703

[122] Panda M, Patra M R. Network intrusion detection using Naïve Bayes. International journal of computer science and network security, 2007, 7(12): 258- 263

[123] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009, 11(1):10-18

[124] Kruegel C, Mutz D, Robertson W, Valeur F. Bayesian Event Classification for Intrusion Detection//Proceedings of IEEE Computer Security Applications Conference. Las Vegas, USA, 2003:14-23

[125] Berral JL, Poggi N, Alonso J, et al. Adaptive distributed mechanism against flooding network attacks based on machine learning// Proceeding of ACM Workshop on Security and Artificial Intelligence. Alexandria, USA, 2008: 43-50

[126] Yan G, Lee R, Kent A, et al. Towards a bayesian network game framework for evaluating DDoS attacks and defense//Proceedings of the 2012 ACM conference on Computer and Communications Security. Raleigh, USA, 2012:553-566

[127] Ariu D, Tronci R, Giacinto G. HMMPayl: An intrusion detection system based on Hidden Markov Models. Computers & Security, 2011, 30(4): 221-241

[128] Joshi S S, Phoha V V. Investigating hidden Markov models capabilities in anomaly detection//Proceedings of 43rd ACM Southeast Conference. Kennesaw, USA, 2005: 98–103

[129] Hu W, Liao Y, Vemuri V R. Robust support vector machines for anomaly detection in computer security//Proceedings of International Conference on Machine Learning and Applications. Los Angeles, USA, 2003: 168-174

[130] Cynthia Wagner, Jerome Francois, Radu State, Thomas Engel. Machine learning approach for IP-Flow record anomaly detection//Proceedings of IFIP Networking. Valencia, Spain, 2011: 28-39

[131] Brauckhoff D, Wagner A, May M. FLAME: a flow-level anomaly modeling engine//Proceedings of the 17th USENIX Security Symposium of the Workshop on Cyber Security & Test. San Jose, USA, 2008: 1-6

[132] Kruegel C, Toth T. Using decision trees to improve signature-based intrusion detection. Lecture Notes in Computer Science, 2003, 2820: 173-191

[133] Khamphakdee N, Benjamas N, Saiyod S. Improving Intrusion Detection System based on Snort rules for network probe attack detection// Proceedings of IEEE International Conference on Information and Communication Technology. Nanjing, China, 2014:69-74.

[134] Selim S, Hashem M, Nazmy T M. Hybrid multi-level intrusion detection system. International Journal of Computer Science and Information Security, 2011, 9(5): 23–29

[135] Blowers M, Williams J. Machine learning applied to cyber operations//Proceedings of Network Science and Cybersecurity. Springer New York, 2014: 155-175

[136] Jemili F, Zaghdoud M, Ahmed M B. A framework for an adaptive intrusion detection system using Bayesian network//Proceedings of IEEE International Conference on Intelligence and Security Informatics. New Brunswick, USA, 2007: 66-70

[137] Conti M, Mancini L V, Spolaor R, et al. Analyzing android encrypted network traffic to identify user actions. IEEE Transactions on Information Forensics and Security, 2015, 11(1): 114-125

[138] Caruana G, Li M. A survey of emerging approaches to spam filtering. ACM Computing Surveys, 2008, 44(2): 1-27

[139] Nelson B, Barreno M, Chi F J, et al. Exploiting machine learning to subvert your spam filter//Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. San Francisco, USA, 2008: 1-9

[140] Hao S, Syed N A, Feamster N, et al. Detecting spammers with SNARE: spatio-temporal network-level automatic reputation engine//Proceedings of 18th USENIX Conference on Security Symposium. Montreal, Canada, 2009: 1-17

[141] Sha HZ, Liu QY, Liu WT, et al. Survey on malicious webpage detection research. Chinese Journal of Computers, 2016, 39(3): 529-542 (in Chinese)
(沙泓州, 刘庆云, 柳厅文等. 恶意网页识别研究综述. 计算机学报, 2016, 39(3): 529-542)

[142] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise//Proceedings of the 2nd ACM SIGKDD international conference on Knowledge discovery and data mining. Portland, USA, 1996: 226-231

[143] G. T. Report. Making the web safer, Technical Report. 2014,12,30

[144] Tzermias Z, Sykiotakis G, Polychronakis M, et al. Combining static and dynamic analysis for the detection of malicious documents//Proceedings of the Fourth European Workshop on System Security. Salzburg, Austria, 2011:1-6

[145] Stolfo S J, Wang K, Li W J. Fileprint analysis for malware detection//Proceedings of the 12th ACM Conference on Computer and Communications Security Workshop on Rapid Malcode. Fairfax, USA, 2005: 1-12

[146] Šrndic N, Laskov P. Detection of malicious pdf files based on hierarchical document structure//Proceedings of the 20th Annual Network & Distributed System Security Symposium 2013. San Diego, USA, 2013: 1-16

[147] Šrndić N, Laskov P. Hidost: a static machine-learning-based detector of malicious files. Eurasip Journal on Information Security, 2016, 2016(1): 22

[148] Xu W, Qi Y, Evans D. Automatically evading classifiers//Proceedings of the 20th Network and Distributed System Security Symposium 2016. San Diego, USA, 2016: 1-15

[149] Rndic N, Laskov P. Practical evasion of a learning-based classifier: a case study//Proceedings of IEEE Symposium Security and Privacy. San Jose, USA, 2014: 197-211

[150] Zhang Yuqing, Lv Shaoqing, Fan Dan. Anomaly detection in online social networks. Chinese Journal of Computers, 2015, 38(10):2011-2027 (in Chinese)
(张玉清, 吕少卿, 范丹. 在线社交网络中异常帐号检测方法研究. 计算机学报, 2015, 38(10):2011-2027)

[151] Wang G, Konolige T, Wilson C, et al. You are how you click: clickstream analysis for Sybil detection//Proceedings of the 22rd USENIX Security Symposium. Washington, USA, 2013: 241-256

[152] Freeman DM. Using naive bayes to detect spammy names in social networks// Proceedings of ACM Workshop on Security and Artificial Intelligence. Berlin, Germany, 2013: 3–12

[153] Viswanath B, Bashir M A, Crovela M, et al. Towards detecting anomalous user behavior in online social networks//Proceedings of the 23rd USENIX Security Symposium. San Diego, USA, 2014: 223-238

[154] Egele M, Stringhini G, Kruegel C, et al. Towards detecting compromised accounts on social networks. IEEE Transactions on Dependable & Secure Computing, 2015, 12(2): 91-98

[155] Thomas K,Grier C,Ma J,et al. Design and evaluation of a real-time url spam filtering service//Proceedings of the Symposium on Security and Privacy. Oakland, USA, 2011: 447-462

[156] Amleshwaram A A, Reddy N, Yadav S, et al. CATS: characterizing automation of Twitter spammers//Proceedings of the 5th International Conference on Communication Systems and Networks. Bangalore, India, 2013: 1-10

[157] Raj S B E, Portia A A. Analysis on credit card fraud detection methods///Proceedings of 2011 International Conference on Computer, Communication and Electrical Technology. Tirunelveli, India, 2011: 152-156

[158] Bhattacharyya S, Jha S, Tharakunnel K, et al. Data mining for credit card fraud: a comparative study. Decision Support Systems, 2011, 50(3): 602-613

[159] Sahin Y, Bulkan S, Duman E. A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 2013, 40(15): 5916-5923

[160] Srivastava A, Kundu A, Sural S, et al. Credit card fraud detection using hidden Markov model. IEEE Transactions on Dependable and Secure Computing, 2008, 5(1): 37-48

[161] Pearl L, Steyvers M. Detecting authorship deception: a supervised machine learning approach using author writeprints. Literary and Linguistic Computing, 2012, 27(2): 183-196

[162] Khanna N, Chiu G T C, Allebach J P, et al. Forensic techniques for classifying scanner, computer generated and digital camera images//Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA, 2008: 1653-1656

[163] Khan M N A, Chatwin C R, Young R C. Extracting evidence from filesystem activity using Bayesian networks. International Journal of Forensic Computer Science, 2007, 1: 50-63

[164] Mukkamala S, Sung A H. Identifying significant features for network forensic analysis using artificial intelligent techniques. International Journal of Digital Evidence, 2003, 1(4): 1-17

[165] Palomo E J, North J, Elizondo D, et al. Visualisation of network forensics traffic data with a self-organising map for qualitative features//Proceedings of 2011 international joint conference on Neural networks. San Jose, USA, 2011: 1740-1747

[166] Palomo E J, North J, Elizondo D, et al. Application of growing hierarchical SOM for visualisation of network forensics traffic data. Neural Networks, 2012, 32: 275-284

[167] Dai Yuan. Research on security evaluation indicator system of internet public opinion in China [M. S. dissertation]. Beijing University of Chemical Technology, China, 2008
(戴媛. 我国网络舆情安全评估指标体系研究[硕士学位论文]. 北京化工大学, 北京, 2008)

[168] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Lansdowne, USA, 1998: 194-218

[169] Rajaraman K, Tan A H. Topic detection, tracking, and trend analysis using self-organizing neural networks///Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin Heidelberg, German, 2001: 102-107

[170] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2008, 2(1-2): 1-135

[171] Vinodhini G, Chandrasekaran R M. Sentiment analysis and opinion mining: a survey. International Journal, 2012, 2(6): 282-292

[172] Alsaedi N, Burnap P, Rana O F. A combined classification-clustering framework for identifying disruptive events//Proceedings of ASE SocialCom Conference. Stanford University, USA, 2014: 1-10

[173] Liu H. Internet public opinion hotspot detection and analysis based on K means and SVM algorithm//Proceedings of 2010 International Conference of Information Science and Management Engineering. Shanxi, China, 2010: 257-261

[174] Li N, Wu D D. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision support systems, 2010, 48(2): 354-368

[175] Zeng J, Zhang S, Wu C, et al. Predictive model for internet public opinion//Proceedings of Fourth International Conference on Fuzzy Systems and Knowledge Discovery. Washington, USA, 2007: 7-11

[176] Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems// Proceedings of IEEE Symposium Security and Privacy. San Jose, USA, 2016: 598-617

[177] Adler P, Falk C, Friedler S A, et al. Auditing black-box models for indirect influence//Proceedings of IEEE International Conference on Data Mining. New Orleans, USA, 2017: 1-10

[178] Henelius A, Kai P, Boström H, et al. A peek into the black box: exploring classifiers by randomization. Data Mining & Knowledge Discovery, 2014, 28(5-6): 1503-1529

[179] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, UK, 2012: 1807-1814

[180] Barreno M, Nelson B, Sears R, et al. Can machine learning be secure?//Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. Beijing, China, 2006: 16-25

[181] Barreno M, Nelson B, Joseph A D, et al. The security of machine learning. Machine Learning, 2010, 81(2): 121-148

[182] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks// Proceedings of IEEE Symposium Security and Privacy. San Jose, USA, 2016: 582-597

[183] Huang S, Papernot N, Goodfellow I, et al. Adversarial attacks on neural network policies. arXiv:1702.02284, 2017

[184] Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks. arXiv:1701.04143, 2017

**ZHANG Lei**, born in 1979, Ph.D. candidate. Her research interests include machine learning and cyberspace security.

**CUI Yong**, born in 1976, Ph. D., professor, Ph. D. supervisor. His current research interests include computer network architecture, mobile computation and cyberspace security.

**LIU Jing**, born in 1993, M. S. candidate. Her current research interests include machine learning, network security and privacy preservation.

**JIANG Yong**, born in 1975, Ph. D., professor, Ph. D. supervisor. His current research interests include computer network architecture, next generation Internet, and mobile computation.

**WU Jian-Ping**, born in 1953, Ph. D., professor, Ph. D. supervisor. His current research interests mainly include computer network architecture, next generation Internet and cyberspace security.

**Background**

Cyberspace security has become an important research area because it affects the development of national security and the stability of society. With the rapid development of new technologies, such as cloud computing, Internet of Thing, big

data, cyberspace security is being confronted with a series of new threats and challenges. There has been much work done in both machine learning and cyberspace security in recent years. A certain research achievements have been gained by academia and industry. However, a lot of security issues based on machine learning still have not been addressed well.

This paper provides an overall review of the application of machine learning for cyberspace security. The authors first summarize the workflow of machine learning in security issues. Furthermore， they categorize the research on cyberspace security into three main subjects: system security, network security and application security. The authors discuss the security features and machine learning algorithms in each of the subjects followed by discussion and comparison on these techniques, and point out the possible research trends in the future.

This work is supported by the National Natural Science Foundation of China (Grant No.61422206).