人工智能在信息安全领域应用的理论基础探究

司德成
长三角（嘉兴）城乡建设设计集团有限公司　浙江　嘉兴　314000

摘　要：随着信息技术迅猛发展，信息安全碰上愈发复杂多样的威胁。人工智能技术靠强大数据处理、模式识别能力，慢慢成为提升信息安全防护水平的关键手段。本文探究人工智能用于信息安全领域的理论基础，讲讲人工智能、信息安全相关概念和特点，分析机器学习、深度学习这些关键技术在信息安全里的应用原理，说说它们在入侵检测、恶意软件检测、防范网络钓鱼等场景的具体应用，剖析应用时面临的挑战以及未来发展趋势。目的是给进一步推动人工智能在信息安全领域深入应用提供理论支持，助力打造更高效、智能的信息安全防护体系。

关键词：人工智能；信息安全；机器学习；深度学习；应用理论

引言

数字化时代，信息成为重要资产。网络空间里，攻击手段五花八门，信息安全形势严峻起来。传统信息安全防护手段，面对新型复杂安全威胁，渐渐显出局限。人工智能技术兴起，给信息安全领域带来新希望。它能自动处理海量数据，识别异常行为模式，快速响应安全事件，大大提升信息安全防护效率和准确性。研究人工智能在信息安全领域应用的理论基础，对发挥人工智能技术优势、应对信息安全挑战意义重大。

1 人工智能与信息安全相关概念及特点

1.1 人工智能的概念与发展

人工智能是让计算机模拟、延伸和扩展人类智能的技术科学，有着漫长又丰富的发展历程。起初，简易专家系统初次体现出人工智能模仿人类思维解决问题的能力。后来，机器学习技术出现，计算机借此能从数据里学习，自动优化性能。当下，深度学习依靠强大的神经网络架构，在诸多领域取得重大进展。近些年来，计算能力飞速提升，积累起海量数据，算法持续创新，这使得人工智能在图像识别、自然语言处理、智能驾驶等好多领域被广泛应用，把人们的生活和工作方式彻底改变。

1.2 信息安全的内涵与现状

信息安全旨在给数据处理系统打造全面的技术与管理安全保护体系，关键是要保证计算机硬件、软件和数据，不被偶然状况或者恶意攻击破坏、更改、泄露。现在，信息安全形势相当严峻。网络攻击手段变得越来越复杂多样，勒索病毒加密用户数据来索要赎金，严重干扰个人和企业正常运作；高级持续威胁（APT）隐蔽地长期潜伏在目标系统，窃取重要信息。物联网、云计算等新兴技术不断普及，让信息安全边界大幅拓展。物联网里大量设备接入网络，安全防护差，容易成为攻击的突破口；云计算环境中数据集中存储和处理，一旦遭到攻击，影响范围特别广。这些情况都大大增加安全风险。

1.3 人工智能应用于信息安全的独特优势

面对海量安全数据，人工智能凭借超强计算能力和高效算法，能快速处理分析，从繁杂数据里挖掘出潜在安全威胁模式，人工处理可做不到这点。人工智能有很强的自适应能力，能实时察觉网络环境变化、攻击手段演变，迅速调整防护策略，实现动态安全防护[1]。它的自动学习能力可以从新安全事件和数据中吸取经验，持续优化安全模型，大幅提升检测、防范各类安全威胁的准确性，给信息安全防护筑牢根基。

2 人工智能关键技术在信息安全中的应用原理

2.1 机器学习技术

监督学习依据已标注的训练数据构建模型，再用模型对新数据实施分类或预测，在信息安全领域常被应用于入侵检测。将大量已知的正常网络行为数据以及入侵行为数据收集起来当作训练集，以此训练分类模型。一旦有新的网络流量数据输入，模型便能判断该数据是否属于入侵行为。诸如支持向量机（SVM）、决策树等常见的监督学习算法，在入侵检测系统中被广泛运用。无监督学习负责处理未标注的数据，其目标是挖掘出数据里的内在结构与模式。在恶意软件检测方面，无监督学习能够借助对大量软件行为特征的分析，找出异常行为模式，进而识别出潜在的恶意软件。运用聚类算法把行为相似的软件归为一类，针对偏离正常聚类的软件再做进一步检测，以判断其是否为恶意软件。半监督学习把少量标注数据和大量未标注数据相结合来开展模型训练。在信息安全领域，要是获取大量标注的安全数据成本比较高，半监督学习的优势就会十分明显。

2.2 深度学习技术

在图像识别、语音识别等领域收获巨大成功，在信息安全方面同样作用重大。卷积神经网络（CNN）能用于恶意软件的图像化分析。把恶意软件的二进制代码转化成图像形式，凭借 CNN 强大的特征提取能力，学习恶意软件的图像特征，实现对恶意软件的准确识别。循环神经网络（RNN）及其变体长短时记忆网络（LSTM）很适合处理序列数据。网络流量数据有时间序列特性，在网络流量分析中，RNN 和 LSTM 可以学习网络流量随时间的变化模式，检测出异常流量变化，像 DDoS 攻击前期流量的异常增长等情况，有效防范网络攻击[2]。生成对抗网络（GAN）由生成器和判别器构成。在信息安全里，GAN 能生成虚假的网络流量或文件，迷惑攻击者，加大攻击者识别真实目标的难度。还能用于检测异常数据，通过训练生成器生成正常数据的分布，判别器判断输入数据是否符合这个分布，从而检测出异常数据，应用于入侵检测等场景。

3 人工智能在信息安全具体场景中的应用

3.1 入侵检测与防御

基于人工智能的入侵检测系统一般由数据采集模块、数据预处理模块、特征提取模块、人工智能模型训练与检测模块和响应模块构成。数据采集模块收集网络流量、系统日志之类的数据，数据预处理模块对数据开展清洗、去噪等操作，特征提取模块提取出有用特征，随后把特征数据输入已训练好的人工智能模型进行检测。一旦检测到存在入侵行为，响应模块即刻采取诸如阻断连接、发出警报等相应防御措施。不少企业运用基于机器学习的入侵检测系统。比如某大型互联网公司借助随机森林算法来训练入侵检测模型，通过采集过去一段时间内的网络流量数据，涵盖正常流量以及已知的入侵流量，以此对模型进行训练。在实际运行过程中，该系统可以精准检测出端口扫描、SQL注入等多种入侵行为，极大提升了网络的安全性，降低了因入侵行为而造成的经济损失。深度学习在入侵检测时能够自动学习复杂特征，对于未知类型的入侵检测有着较高的准确率。深度学习模型训练需要耗费大量的计算资源和时间，且模型的可解释性欠佳，很难明白其决策过程，在一定程度上对其在一些对安全性和可解释性要求比较高的场景中的应用形成了限制。

3.2 恶意软件检测

传统恶意软件检测依靠特征匹配，提取代码特征与病毒库比对来识别。对已知软件检测率高，新变种常无效，因特征码不同且库更新慢。人工智能从行为、结构等多维度检测恶意软件。以深度学习中RNN给恶意软件行为序列建模，学其系统操作模式，像文件读写、网络连接等，能有效检出新型变种。再结合机器学习集成法，融合多模型结果，提升检测准确性。

3.3 网络钓鱼防范

网络钓鱼常发伪装成合法机构的邮件或短信，骗用户交出账号密码、信用卡信息这类敏感内容。手段不停变花样，运用社会工程学技巧，结合时事热点弄出欺骗内容，特别容易让人上当。网络钓鱼让个人和企业遭受严重经济损失，还可能致使个人隐私泄露、企业声誉变差[3]。人工智能靠分析邮件文本内容、发件人信息、邮件结构等诸多特征识别钓鱼邮件。自然语言处理技术能对邮件文本做情感分析、提取关键词，以此判断邮件内容有没有欺诈企图。机器学习算法可学习正常与钓鱼邮件的特征模式，构建分类模型，快速又准确地给新收到的邮件分类。就是用朴素贝叶斯算法分类邮件，训练模型学习大量正常及钓鱼邮件特征，实际用起来能有效识别网络钓鱼邮件。

4 人工智能在信息安全领域应用面临的挑战与对策

4.1 数据质量与隐私问题

信息安全领域里，数据会有噪声、不完整、标注不准等状况。像入侵检测数据，有误报数据被错标成真实入侵数据的情形，这会让训练出的模型准确率降低，造成误判。可见，数据质量对人工智能模型在信息安全应用中的性能影响很大。收集、使用安全数据时，会涉及大量用户隐私信息。拿恶意软件检测来说，可能要获取用户设备上软件行为数据，而这些数据含有用户隐私。怎样在借数据提升信息安全防护能力的同时，保护好用户数据隐私，是人工智能在信息安全领域应用面临的一大难题。面对数据质量问题，能运用数据清洗、数据增强等技术。数据清洗把噪声和错误标注数据去掉，数据增强则通过对现有数据做变换，生成更多训练数据，提升数据多样性和质量。至于数据隐私保护，可用加密技术给数据加密，在数据使用时采用联邦学习等技术，做到数据“可用不可见”，在保护数据隐私的基础上开展模型训练。

4.2 模型安全性与可靠性

对抗攻击是攻击者给输入数据加上精心设计的微小干扰，让人工智能模型输出错误结果。在信息安全领域，恶意攻击者会用对抗攻击手段骗入侵检测系统、恶意软件检测系统等。比如攻击者给正常网络流量数据加微小干扰，使其绕开入侵检测系统检测，严重威胁网络安全[4]。信息安全应用中，模型可靠性十分关键。不可靠的模型会造成误报或漏报，影响信息安全防护效果。

模型可解释性也很重要，尤其在关键决策场景。像金融领域信息安全防护，监管部门可能要求解释模型决策过程，保证安全防护措施合理合规。为提高模型安全性，能用对抗训练技术，让模型训练时学习对抗攻击特征，增强鲁棒性。对于模型可靠性，可采用多模型融合方法，综合多个模型结果，提升决策准确性。在模型可解释性上，能用解释性人工智能技术，如局部可解释的模型无关解释（LIME）、SHapley可加性解释（SHAP）等，将模型决策过程可视化解释，提高模型可信度。

4.3 人才短缺与技术更新

人工智能用于信息安全领域，需要既懂信息安全，又熟悉人工智能技术的复合型人才。可当下，这类交叉领域人才短缺得厉害。一方面，信息安全专业人才对人工智能技术掌握不够；另一方面，人工智能专业人才对信息安全知识了解有限，没法满足实际应用需要。信息安全和人工智能技术都在快速发展，新攻击手段、人工智能算法不断冒出来[5]。这就要求从业人员不断更新知识，跟上技术发展节奏。就拿量子计算技术来说，它的发展让传统加密算法有被破解风险，所以得研究量子抗性加密算法和信息安全防护技术，这对人才知识更新能力提出了更高要求。高校和培训机构加强信息安全与人工智能交叉学科建设，开设相关专业课程，培养复合型人才。企业也要加强员工培训，提供技术交流、学习机会，鼓励员工自主学习，持续更新知识。另外，行业得建立技术跟踪机制，及时掌握最新技术发展动态，促进技术交流合作，一起推动人工智能在信息安全领域的应用发展。

结语

信息安全领域里，人工智能的应用潜力尽显，为应对日趋复杂的信息安全威胁，开拓出新思路与办法。探究人工智能理论基础后，可知其相关技术于信息安全里的应用原理、具体场景，还有应用时面临的挑战及应对之策。当下虽有数据质量、模型安全、人才匮乏等状况，不过随着技术持续发展、趋于完善，这些难题会渐渐解决。未来，人工智能与信息安全技术有望深度交融，进一步提升信息安全防护的智能化程度，给数字化社会的平稳发展筑牢根基。对人工智能在信息安全领域的发展态势，我们需持续留意，强化技术创新，注重人才培育，充分挖掘人工智能技术优势，携手打造更安全可靠的信息网络环境。

参考文献

[1]张少昌.人工智能安全框架下信息安全防护机制的探索[J].中国信息界,2025,(01):149-151.

[2]覃露.基于人工智能的信息安全防护系统研究[C]//中国智慧工程研究会.文化传承与现代化治理学术交流会论文集.广西民族大学;,2024:316-317.

[3]蒋红亮.基于人工智能的互联网信息安全风险评估与防控策略研究[J].中国新通信,2024,26(20):32-34+172.

[4]伍星宇,裴旭斌,陈泽堃.基于机器学习的信息安全技术研究[J].无线互联科技,2020,17(20):57-60+65.

[5]魏敏.基于人工智能的计算机网络信息安全防护模式研究[J].信息记录材料,2024,25(11):130-132.