AI in Security
人工智能赋能安全
应用案例集

2021 年 2 月

目录

前言

当前，新一轮科技革命和产业变革正在萌发，在 5G、大数据、云计算、深度学习等新技术的共同驱动下，人工智能（AI）作为新型基础设施的重要战略性技术加速发展，并与社会各行各业创新融合，引发链式变革。特别是在网络空间安全防护领域，人工智能在威胁识别、态势感知、风险评分、恶意检测、不良信息治理、骚扰诈骗电话检测、灰黑产识别等方面有其独特的价值和优势，应用需求呈现跨越式发展，产生了显著的溢出效应。为探索解决行业安全应用前沿问题，打造 AI in Security（人工智能赋能安全）最佳实践“样板间”，并推动全球信息通信行业对人工智能在在安全领域形成共识，形成普及应用和规模发展，提升网络空间智能安全防护水平，特编制本案例集。

本案例集以“人工智能赋能安全，共筑智慧网络新未来”为核心思想，首先从技术发展、战略布局、标准制定等方面，阐释了人工智能技术和产业的最新发展；其次，围绕网络空间安全新特点和人工智能赋能新优势两个维度，分析了人工智能赋能安全领域的发展蓝海；在此基础上，结合行业领先企业的最佳实践，分为通信网络安全、内容安全、数据安全、业务安全、终端安全五个篇章，详细介绍了人工智能赋能网络空间安全的具体应用模式和工作案例；最后，站在全球网络空间安全防护的战略高度，提出了未来展望，倡议产业各方开放合作，积极推动人工智能在网络安全产业中发挥更大价值，共同构建人工智能赋能安全共赢生态。

1 人工智能技术发展现状与趋势

1.1 全球人工智能技术与产业发展

人工智能作为研究开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的一门技术科学，通过对数据的采集、分析和挖掘，形成有价值的信息和知识模型，实现了对人类智能行为的模拟，具备不同环境下的自适应特性和学习能力。其中，人工智能一般包括知识、数据、模型（算法）、算力、应用场景等要素，并涉及机器学习、知识图谱、语音识别、自然语言处理、计算机视觉、生物特征识别等关键技术。

近年来人工智能的繁荣得益于三个主要驱动力。首先特征降维、人工神经网络（ANN）、概率图形模型、强化学习和元学习等方面的新理论和新技术层出不穷，在学术和工业领域都取得明显突破。第二，计算能力的进步使许多计算资源消耗型机器学习算法可以大规模普及。第三，在大数据时代，数据资源的极大丰富可以让机器学习模型泛化能力更强。尤其是深度学习技术使我们能够从更多数据中构建合理的人工智能模型，让机器发挥更大的潜力，也让各种任务取得更好的结果。深度学习极大地改变了人们的生活，并重塑了传统的人工智能技术，人工智能理论建模、技术创新、软硬件发展等各方面要素整体推进。

特别是随着 5G 时代的全面到来，在数字化生活方面，“高速连接+感官智能”将催生人机交互新应用，视觉、听觉、触觉智能会在个人穿戴、家居设备中快速渗透，展现丰富多彩的智慧生活；在数字化生产方

面，“可靠连接+专用智能”将催生智能制造新业态。凭借高可靠、低时延特征，5G 将整合工业生产各领域离散网络，推动人工智能深度融入制造业全流程、全环节，大幅提升传统产业的生产效率；在数字化治理方面，“广域连接+通用智能”将催生智慧治理新模式。

可以预测到，各类人工智能应用可大幅度提升智能制造水平、社会智能水平，推动制造业转型和数字经济建设，推动社会各领域从数字化、网络化向智能化加速跃升，从而深刻改变甚至颠覆现有人类社会的生产生活方式。有数据显示，到 2025 年，全世界人工智能市场规模将超过 6 万亿美元。人工智能产业发展势头良好，与行业融合应用不断深入，发展前景无限可期。

1.2 全球人工智能战略法规布局

当前，全球人工智能蓬勃发展，世界主要发达国家和联盟纷纷布局人工智能发展战略规划，构建人工智能安全相关标准规范体系，人工智能已经成为全球各国战略竞争焦点。

2016 年，美国发布了《美国国家人工智能研究与发展战略规划》，目标是投资研究，开发人工智能协作方法，解决人工智能的安全，道德，法律和社会影响，为人工智能培训创建公共数据集，并通过标准和基准评估人工智能技术，首次将人工智能上升到美国国家战略高度。近期，美国政府启动了“美国人工智能计划”，是前期国家人工智能研发战略的延伸，主要包括研究和开发、释放资源、道德标准、自动化和国际推广等五方面，要求确保人工智能系统安全可靠。

2018 年，欧盟委员会先后发布了《欧盟人工智能》报告，并提出了一项数字欧洲计划，明确了欧盟人工智能行动计划，主要举措包括承诺将欧盟对人工智能的投资从 2017 年的 5 亿欧元增加到 2020 年底的 15 亿欧元，同时建立欧洲人工智能联盟，并设立“人工智能高级别小组”作为其指导小组，负责起草道德准则供成员国审议。同年，欧盟成员国签署了人工智能合作宣言。2020 年，欧盟委员会进一步发布了《关于人工智能、物联网、机器人对安全和责任的影响的报告》。

人工智能也成为日韩谋划新一轮产业升级的重要抓手。2016 年，日本提出了超智能社会 5.0 战略，将人工智能作为实现超智能社会 5.0 的核心。同时，明确提出设立“人工智能战略会议”，通过产学官相结合的战略高度作为实现第四次产业革命的具体措施。韩国第四次工业革命委员会审议通过了人工智能研发战略，主要包括确保人才、技术和基础设施等三个方面。其中，预计在 2022 年前新设 6 所人工智能研究生院，拥有 1370 名人工智能高级人才，培养 350 名人工智能高级研究人员的计划，并投资 20 亿美元用于人工智能研究。

2017 年 7 月，中国正式印发了《新一代人工智能发展规划》，从战略态势、总体要求、资源配置、立法、组织等各个层面阐述了中国人工智能发展规划，要求加强人工智能标准框架体系研究，到 2020 年初步建成人工智能技术标准体系，其中包括人工智能网络安全、隐私保护等技术标准，鼓励各行业各单位参与或主导制定国际标准；同时，工信部在《“十三五”技术标准体系建设方案》中也明确提出，到 2020 年完善人工智能网络安全产业布局，形成人工智能安全防控体系框架，在《促进新一代人工智能产业发展三年行动计划》中提出，以信息技术与制造技术深度融合为主线，推动新一代人工智能关键技术创新与产业化协

同发展。2020 年 7 月 27 日，国家标准化管理委员会、中央网信办、国家发改委、科技部、工信部等五部委印发《国家新一代人工智能标准体系建设指南》，要求到 2021 年，明确人工智能标准化顶层设计，完成安全/伦理等 20 项以上重点标准的预研工作，到 2023 年初步建立人工智能标准体系，全面形成标准引领人工智能产业全面规范化发展的新格局。特别是 2020 年以来，中国积极布局推进以 5G 为引领、人工智能为核心的“新基建”战略。随着 5G 网络建设提速，与云、边、端等基础设施协同，大大降低了人工智能使用门槛，全面推动人工智能技术深度融入经济社会发展。可以说，5G 将使人工智能更泛在，人工智能将使 5G 更智慧。

1.3 人工智能相关标准规范

ISO/IEC、ITU-T、IEEE 等国际标准组织以及各国家/区域标准组织均高度重视了人工智能相关标准规范研究编制工作。

- **ISO/IEC JTC1**

2017 年 10 月，ISO/IEC JTC1 成立人工智能分委员会 SC42，专门负责人工智能标准化工作。SC42 下设 5 个工作组：基础标准（WG1）、大数据（WG2）、可信赖（WG3）、用例与应用（WG4）、人工智能系统计算方法和计算特征工作组（WG5），以及人工智能传播与外联咨询组（AHG1）和智能系统工程咨询组（AG2）等。其中主要标准项目包括：ISO/IEC TR 24027《信息技术人工智能人工智能系统中的偏差与人工智能辅助决策》、TR 24028《信息技术人工智能人工智能可信度概述》、TR 24029-1《人工智能神经网络鲁棒性评估第 1 部分：概述》、AWI 24029-2《人工智能神经网络鲁棒性评估第 2 部分：形式化方法》、CD23894《信息技术人工智能风险管理》和 AWI TR 24368《信息技术人工智能伦理和社会关注概述》等。

- **ITU-T**

ITU-T 一直致力于解决智慧医疗、智能汽车、垃圾内容治理、生物特征识别等人工智能应用中的安全问题。2017 年和 2018 年，ITU-T 均组织了“AI for Good Global”峰会，重点关注确保人工智能技术可信、安全和包容性发展的战略，以及公平获利的权利。ITU-T 中，SG17 安全研究组和 SG16 多媒体研究组均开展了人工智能安全相关标准研制工作，特别是 ITU-T SG17 已经计划开展人工智能赋能安全相关标准化项目的讨论和研究。同时，ITU-TSG17 安全标准工作组下设远程生物特征识别问题组（Q9）和身份管理架构和机制问题组（Q10），主要负责 ITU-T 生物特征识别标准化工作；其中，Q9 关注生物特征数据的隐私保护、可靠性和安全性等方面的各种挑战。

- **IEEE**

IEEE 持续开展多项人工智能伦理道德研究，发布了 IEEE P7000 系列等多项人工智能伦理标准和研究报告，用于规范人工智能系统道德规范问题，包括：IEEE P7000《在系统设计中处理伦理问题的模型过程》、

P7001《自治系统的透明度》、P7002《数据隐私处理》、P7003《算法偏差注意事项》、P7004《儿童和学生数据治理标准》、P7005《透明雇主数据治理标准》、P7006《个人数据人工智能代理标准》、P7007《伦理驱动的机器人和自动化系统的本体标准》、P7008《机器人、智能与自主系统中伦理驱动的助推标准》、P7009《自主和半自主系统的失效安全设计标准》、P7010《合乎伦理的人工智能与自主系统的福祉度量标准》、P7011《新闻信源识别和评级过程标准》、P7012《机器可读个人隐私条款标准》、P7013《人脸自动分析技术的收录与应用标准》等。

- **美国 NIST**

美国国家标准与技术研究院（NIST）专注于理解人工智能可信度的研究，并将这些指标纳入未来的标准，也建议在监管或采购中引用的人工智能标准保持灵活性，以适应人工智能技术的快速发展；制定度量标准以评估人工智能系统的可信赖属性；研究告知风险、监控和缓解风险等人工智能风险管理;研究对人工智能的设计、开发和使用的信任需求和方法；通过人工智能挑战问题和测试平台促进创造性的问题解决等。2019 年 8 月，NIST 发布了《关于政府如何制定人工智能技术标准和相关工具的指导意见》，概述了多项有助于美国政府推动负责任地使用人工智能的举措，并列出了一些指导原则，这些原则将为未来的技术标准提供指导。

- **欧盟**

2019 年 4 月 9 日，欧盟委员会（EC）任命的人工智能高级专家小组发布人工智能道德准则，提出了“可信任人工智能”应当满足的 7 个原则：（1）人类的力量和监督；（2）技术的可靠性和安全性；（3）隐私和数据管理；（4）透明性；（5）多样性、非歧视性和公平性；（6）社会和环境福祉；（7）可追责性。

下一阶段，欧盟委员会将启动人工智能道德准则的试行，邀请工业界、研究机构和政府机构对该准则进行测试和补充。

- **GSMA**

2019 年 6 月 27 日，GSMA 联合 11 家产业伙伴宣布成立 AI in Network 特别工作组，研究人工智能在移动网络的关键应用，共同构筑智能自治网络时代。四个月后，特别工作组完成发布了《AI in Network 智能自治网络案例报告》白皮书，该报告集中展示了人工智能技术应用于移动网络的规划、部署、维护、监控、优化、节能和安全防护方面的案例。

2020 年 7 月 2 日， GSMA 联合 12 家产业伙伴宣布成立 AI in Security 特别工作组，研究人工智能在安全领域的关键应用，共同构建智能网络安全时代。

- **中国 TC260、CCSA**

中国国家标准化管理委员会于 2018 年 1 月正式成立国家人工智能标准化总体组，承担人工智能标准化工作的统筹协调和规划布局，负责开展人工智能国际国内标准化工作，目前已发布《人工智能安全标准化白皮书（2019 版）》、《人工智能伦理风险分析报告》等，正在研究人工智能术语、人工智能伦理风险评估等标准。

中国全国信息安全标准化技术委员会（TC260）的人工智能安全相关标准主要集中在生物特征识别、智慧家居等人工智能赋能安全领域，以及与数据安全、个人信息保护相关的支撑领域。主要包括：基础共性标准方面有《人工智能安全标准研究》、《人工智能应用安全指南》等；生物特征识别安全标准方面有 GB/T 20979-2019《信息安全技术 虹膜识别系统技术要求》、GB/T 36651-2018《信息安全技术 基于可信环境的生物特征识别身份鉴别协议》、GB/T 37076-2018《信息安全技术 指纹识别系统技术要求》、GB/T 38671-2020《信息安全技术 远程人脸识别系统技术要求》，在研标准《信息安全技术 生物特征信息保护》；智慧家居安全标准方面有《信息安全技术 智能家居安全通用技术要求》、《信息安全技术 智能门锁安全技术要求和测试评价方法》等在研标准；数据安全和个人信息保护标准方面有 GB/T 35273-2020《信息安全技术 个人信息安全规范》、GB/T 37964-2019《信息安全技术 个人信息去标识化指南》、GB/T35274-2017《信息安全技术 大数据服务安全能力要求》、GB/T 37932-2019《信息安全技术 数据交易服务安全要求》、GB/T 37988-2019《信息安全技术 数据安全能力成熟度模型》等。中国通信标准化协会（CCSA）以人工智能在具体应用场景为主，已开展汽车电子、智能家居等方面标准研究工作，目前已发布 YDB 201-2018《智能家居终端设备安全能力技术要求》、T/CSHIA 001-2018《智能家居网络系统安全技术要求》、T/CCSA 284-2020《智能家居终端设备安全能力测试方法》等标准；在研标准包括《人工智能产品、应用及服务安全评估指南》、《人工智能终端产品标准体系研究》、《移动智能终端人工智能能力及应用个人信息保护技术要求及评估方法》等。

2 人工智能赋能安全内涵与意义

毫无疑问，当前网络空间已经进入到人工智能时代。人工智能对网络空间产生了深远影响，使人工智能时代的安全问题呈现出新的趋势，有了新的动向。一是攻击者开始运用人工智能发起新型网络攻击，例如基于人工智能的高级持久威胁；二是出现了针对人工智能系统本身的攻击或欺骗，从而导致分类或预测结果不正确；三是人工智能开始赋能安全，也可称为人工智能保障安全，是指基于相似或先前的活动而利用人工智能来自动识别和/或响应潜在网络威胁（包括前两个动向中的新威胁）的工具和技术。

2.1 人工智能时代网络空间安全新特点

如上所述，网络安全领域不断涌现出与人工智能相关的新应用，例如恶意代码分析、恶意域名检测、异常检测、入侵检测、恶意流量识别、手机恶意软件检测和网络钓鱼防护等。在人工智能赋能安全蓬勃发展浪潮中，机器学习技术在应对网络空间威胁方面起着至关重要的作用。如果特征提取非常准确，无论是

否采用深度学习模型，机器学习都可以有效建模。然而特征提取并非易事，尤其是基于机器学习的网络安全模型。例如，为了使机器学习模型能够识别恶意软件，我们必须手动编排与恶意软件相关的各种功能，这无疑限制了威胁检测的效率和准确性。由于机器学习算法根据预定义的特定功能工作，意味着没有预定义的特征将逃避检测，因此可以得出结论，大多数机器学习算法的性能取决于特征提取的准确性。

鉴于传统机器学习的明显缺点，深度神经网络（也称为深度学习）成为新的研究热点。传统机器学习和深度学习之间在概念上的巨大差异在于，深度学习可用于直接训练原始数据而无需手动提取特征。深度学习可以发现数据之间的非线性相关性。由于具有很强的泛化能力，深度学习模型可以支持新文件类型和未知攻击的检测，这在网络安全防御中是非常明显的优势。近年来，深度学习在防止网络安全威胁，特别是在防止 APT 攻击方面取得了长足的进步。很多研究成果表明，深度神经网络可以学习 APT 攻击的高级抽象特征，即使它们采用了最先进的规避技术。

人工智能给我们带来另一个好处是，随着网络攻击的数量和复杂性不断增长，人工智能正在帮助安全运营分析师提前发现威胁。通过从数以百万计的研究论文，博客和新闻报道中收集威胁情报，人工智能可以辅助提供决策，帮助安全分析师应对每日数以千计的警报，从而大大减少响应时间。

从方法论角度看，与保护系统的防御者相比，攻击者发起攻击要容易得多。攻击者可以选择攻击的时间，通常对目标系统机型预先深入了解，为发起精准攻击，攻击者有时会准备数年或数月的时间，并且可以相当迅速地部署新攻击。这就是所谓的不对称问题，它是网络安全中的核心问题。防御者必须经常面对未知的对手（其攻击方法和时机也未知）的所有攻击，提供持续，高强度的防御。防御者通常还需要在开发防御技术与实际部署防御技术之间存在较长的滞后时间，并且通常缺乏评估其绩效的指标。另一方面，防御者也有一些手段来化解这种不对称。例如，更改系统配置以让攻击者投入更多时间和精力。人工智能可以通过解决操作、透明度和数据访问权限等方面的不对称性来提升防御者能力。在未来，人工智能工具可能会代替网络安全防御者，通过承担枯燥、重复和高吞吐量的任务来减轻运营负担。通过维持稳定有效的防护体系让人工智能成为网络安全防御者的有益补充。此外，人工智能可以评估网络防御的强度，预测攻击并评估攻击的影响。

2.2 人工智能时代网络空间安全新挑战

近年来人工智能给网络安全带来巨大挑战。从网络技术到业务应用，人工智能在不同层面给网络空间带来了前所未有的变化，进而改变了网络空间的生态系统。

在网络技术方面，基于机器学习的僵尸攻击者能够重复进行繁琐的任务，这对于人类攻击者是不可想象的。总体而言，僵尸网络攻击的机器学习算法变得越来越复杂和准确。例如，僵尸系统进行垃圾邮件攻击时，可以自我更新让下次攻击变得更加巧妙。网络防御者需要以更具创新性的解决方案做出反击。有时僵尸网络攻击者会配合采用基于社会工程学方法的 APT 攻击，带来巨大的社会问题。根据来自不同国家的报告，近年来，由人工智能控制的网络攻击非常普遍。例如，早在 2018 年就发生了人工智能配合网络攻击的安全事件，TaskRabbit（自由职业者及其客户的在线市场）遭到黑客攻击。黑客从用户数据中发现了 375

万该网站的用户的社会安全号码和银行账户详细信息，该事件属于严重的用户隐私泄露事件。攻击的发起是由人工智能控制的大型僵尸网络完成的，该僵尸网络使用大量计算机对 TaskRabbit 的服务器进行大规模的 DDoS 攻击。另一个例子是 Instagram 遭受到两次网络攻击。从 2019 年 8 月开始，这家社交媒体巨头的用户发现他们的账户信息已被黑客更改，从而无法登录。2019 年 11 月，Instagram 代码中的 bug 导致数据泄露，该数据泄露在用户浏览器的 URL 中显示用户的密码，可以确定这是一个巨大的安全问题。据推测，攻击者采用基于人工智能的工具来扫描 Instagram 服务器漏洞。

在应用层，比较受关注的领域是使用人工智能生成和伪造具有欺骗性的信息。例如，人工智能假装是真实的用户，模仿人类的行为进行互联网操作，而这些行为很难通过传统方法和真实的用户行为加以区分。社交网络中经过训练的机器人可以自动生成和传播虚假新闻已经在几年前开始出现。它们可能通过网络暴力行为获取利益。在很多国家，很多这类行为的目标是对议员或总统选举等重大事件施加影响。最近一种基于生成对抗网络（GAN）的新型黑客机器人——“Deep fake”震惊了整个互联网世界。Deep fake 是完全虚假的数字创作（文本，音频，图像或视频），这些多媒体信息是通过基于机器学习的现有数据采样生成的。为了说明 Deep fake，我们可以和对抗样本之间进行比较，后者是旨在对机器感知系统本身施加影响的输入，而 Deep fake 的目标是欺骗人。遗憾的是对 Deep fake 监测方法的研究还非常不足，很难进行实用。目前也有些方法通过人工发现来检测 Deep fake，但技术进步可能很快会增加人工监测的难度，甚至让 Deep fake 多媒体信息变得肉眼难辨。

除了上述人工智能带来的网络空间威胁外，人工智能系统本身也可能受到攻击或欺骗，从而导致错误的分类或预测结果。例如，在对抗性环境中，对训练样本的修改将导致对人工智能模型的攻击，而测试样本修改将导致所谓的“逃避攻击”。对抗环境中的攻击旨在破坏各种人工智能应用程序的完整性和可用性，并通过采用对抗性样本误导神经网络，从而导致分类器得出错误的分类，这就需要通过相应的防御措施来应对对抗攻击。随着机器学习模型变得越来越复杂，数据集越来越大，集中式训练方法已无法适应这些新要求。诸如 Google 发起的联合学习之类的分布式学习模式已经出现，使许多智能终端能够以协作方式学习共享模型。但是，所有训练数据都存储在终端设备中，带来了许多安全挑战。如何确保模型不被恶意窃取，以及如何构建具有隐私保护功能的分布式机器学习系统是一个主要的研究热点。

2.3 人工智能赋能安全的新需求

为了最大程度地提升网络空间安全水平，亟需提出智能化、创新性的网络安全防御方法，以高效应对日益复杂、花样频多的风险和威胁。在这一背景下，人工智能应用于安全、人工智能赋能安全、乃至人工智能重塑安全已经成为大势所趋。网络空间在应对安全挑战时对人工智能技术提出了迫切需求，即在获取历史记录和当前安全状态数据后，通过人工智能建模做出智能决策，可以为网络空间提供自适应的安全防护。具体可概括为如下三个方面：

（1）提升复杂数据分析能力。网络安全、业务安全、信息安全等领域，涉及到病毒、攻击报文、图片、文本等信息呈现出爆炸式增长趋势。传统的安全特征分析技术难对海量安全信息进行分析处理，需要引入人工智能辅助分析。

（2）提升自适应防护能力。发现网络攻击、分析提炼监测规则、制定实施安全规则，是安全监测能力提升需要的一个周期。在此周期中，依靠人工分析、制定、实施规则，难以及时针对安全态势做出自适应调整，导致在日益复杂的安全环境中实现主动防御成为巨大挑战，需要引入人工智能进行自适应防护。

（3）辅助降低专业技能要求。网络攻击的复杂度日益增加，样本数量日益增长，对相关人员的专业技能构成了严峻的挑战。纯粹依靠人工的方式对日趋复杂的攻击行为进行分析，已成为不可能完成的任务，需要引入人工智能分析降低专业依赖度。

因此，人工智能赋能安全，可以综合判断并选取最优策略，高效、精准、自适应提升安全风险和威胁的监测、预警和处置等全流程工作效率，这与网络空间安全积极防御体系是逻辑自洽的，可以预见，网络空间安全天然是人工智能技术大显身手的重要发力点。

3 人工智能赋能安全应用案例

3.1 通信网络安全篇

3.1.1 身份认证

基于零信任架构的身份认证

【场景描述】

目前，绝大多数企业都还是采用传统的网络分区和隔离的安全模型，用边界防护设备划分出企业内网和外网，并以此构建企业安全体系。在传统的安全体系下，内网用户默认享有较高的网络权限，而外网用户如异地办公员工、分支机构接入企业内网都需要通过 VPN。不可否认传统的网络安全架构在过去发挥了积极的作用，但是在高级网络攻击肆虐，内部恶意事件频发的今天，传统的网络安全架构需要迭代升级。

随着移动互联网、物联网及行业应用的爆发式增长，未来移动通信将面临千倍数据流量增长和千亿设备联网需求。5G 作为新一代移动通信技术发展的方向，将在提升移动互联网用户业务体验的基础上，进一步满足未来物联网应用的海量需求，与工业、医疗、交通等行业深度融合，实现真正的“万物互联”， 5G 网络面临大量新增的 IOT 设备及其可穿戴设备，传统的用户管理机制在开户，认证等方面成本高昂，已经不能完全满足 5G 用户管理的需求，另外，5G 支持多种接入技术(如 4G 接入、WLAN 接入以及 5G 接入)，垂直行业的设备和网络使用其特有的接入技术，目前不同的接入网络使用不同的接入认证方式，为了使用户可以在不同接入网间实现无缝切换，5G 网络亟需采用一种统一的认证框架，实现灵活并且高效地支持各种应用场景下的双向身份鉴权。

【技术方案】

一、技术方案概述

零信任的核心思想可以概括为：网络边界内外的任何访问主体（人/设备/应用），在未经过验证前都不予信任，需要基于持续的验证和授权建立动态访问信任，其本质是以身份为中心进行访问控制。面对内外部网络攻击的数量和复杂性不断增加，通过人工智能持续学习、自我进化能力实现无特征的检测，做到真正洞察威胁本质，更有效的阻止未知风险，利用深度学习算法训练数以万计的数据，利用多层次的检测技术，应用高检出率和低误报率的算法模型，并借助海量数据，使用特征训练不断完善算法模型，与此同时，辅助信誉机制，行为分析和基因特征等技术，构建起完善的网络防御体系。

本方案整体架构基于零信任思想共包括三大能力：统一身份管理系统、零信任驱动认证机制、可信接入访问网关。其中统一身份管理系统是零信任架构的基础能力，包括用户身份的集中管理、设备身份的集中管理以及基础权限构建模型。基于统一身份管理系统，可以通过在可信的访问主体和可信的访问客体之间建立可信的基础权限，来实现对资产设备的可信化访问。零信任驱动认证机制是整个方案架构的核心能力，首先基于用户对设备的访问行为构建出用户和设备的行为信用基线库，再依托持续信任积分评估模型，对访问主体的全部访问过程进行智能化行为分析，对访问客体的风险系数进行智能化调整，进而实现对用户和资产设备的可信度进行持续的信任评估，根据评估结果动态调整访问控制策略，最后通过可信接入访问网关实现对资产设备的动态访问权限控制。

零信任架构的支撑系统称为控制平面，其他部分都称为数据平面，数据平面由控制平面指挥和配置，访问受保护的访问客体首先需要经过控制平面处理，包括对用户和设备的身份认证与授权，如果用户需要访问安全等级更高的设备，那么需要执行更高强度的认证。该方案在架构实现上将控制平面和数据平面进行了分离，其中统一身份管理系统和可信接入访问网关承载了数据平面的基础数据和策略部署，零信任驱动认证机制则主要充当控制平面，接收来自动态信任评估的用户信任积分评估和设备风险评估结果，并动态调整访问控制策略，进而下发到数据平面策略执行点执行访问控制。

本方案整体系统架构如下图所示：

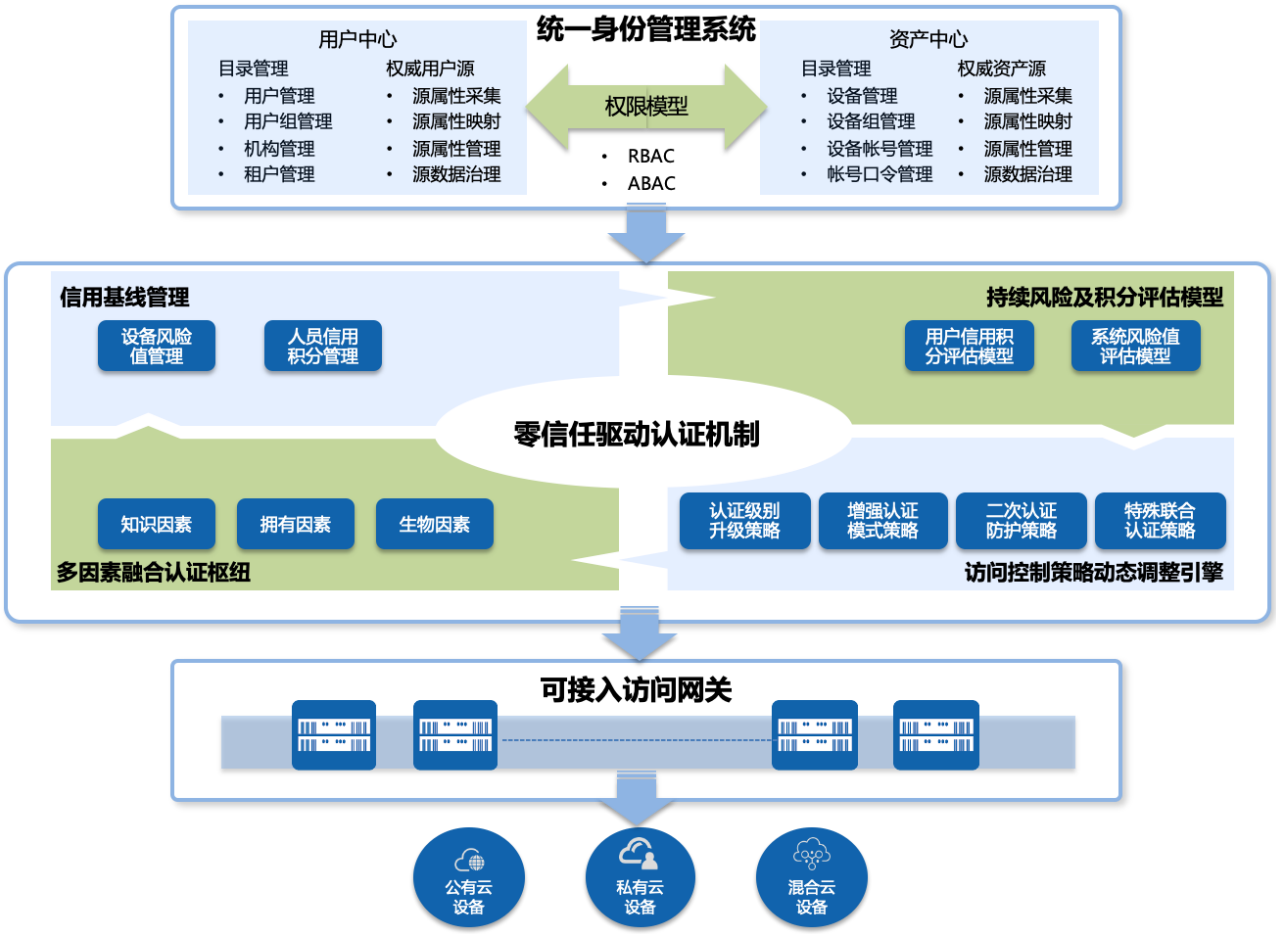

图 1 整体系统架构图

二、关键技术

该方案充分利用三大安全能力，构建基于可信用户和可信设备的访问主体，并对访问行为进行持续信任评估和动态策略调整，以达到无边界的最小权限访问控制。关键技术包括：用户中心及资产中心、信用基线管理、持续风险及信任积分评估模型、访问控制策略动态调整引擎、多因素融合认证枢纽和可接入访问控制网关。

1. 用户中心及资产中心

零信任的本质是以身份为中心，基于身份的治理，构建企业身份边界，基于用户和设备之间的权限认证，重建信任。用户中心是为企业构建的一套用户身份集中化统一管理中心，主要解决企业信息化建设过程中由于业务系统繁多导致的人员账号孤立维护、信息孤岛严重、缺乏身份唯一性、数据难以整合，造成安全运维和管理工作困难的问题。用户中心通过构建统一的用户数据模型，集中梳理企业组织架构信息及用户全生命周期管理机制，对企业各信息系统用户数据进行汇聚，实现多身份源的整合，形成一套权威统一数据源，并对数据进行统一治理，提供用户数据对外共享消费能力。

资产中心是为企业构建的一套资产集中化统一管理中心，主要解决企业信息化建设过程中由于业务系统繁多导致的资产分散、信息不全面、维护属性差异大、未知资产存在安全隐患、数据难以整合，造成安

全运维和管理工作困难的问题。用户中心通过构建统一的资产数据模型，集中梳理企业资产信息及资产全生命周期管理机制，对企业各业务系统内的资产数据进行汇聚，实现多资产源的整合，形成一套权威统一数据源，并对数据进行统一治理，提供资产数据对外共享消费能力。

2. 信用基线管理

零信任得以实现的基础是身份的识别，而身份识别的基础是信任，信用基线管理围绕对用户身份的信任和对设备的信任两个主要纬度进行基础信息构建，包括资产设备风险值信息和人员信用积分信息。用户访问设备的鉴权过程依赖动态风险值和信任积分的评估结果，且评估行为是持续的，伴随整个访问过程。一旦访问过程发生行为异常或环境异常，就下发至访问控制策略动态引擎，自动调整访问权限，保证业务访问的最小权限，确保用户-设备的合法，从而确保访问主体的行为合法，同时，评估结果又会回转到信用基线管理，保证用户和设备的身份识别是动态可变的。

3. 持续风险及信任积分评估模型

持续风险及信任积分评估模型在零信任整体架构中起到承上启下的核心环节，向上与信用基线管理模块对接，信用基线管理模块为持续风险及信任积分评估模型提供访问主体在访问客体上访问行为基础分析结果；向下则为访问控制策略动态调整引擎输入动态授权的依据，通过对资产设备的持续风险评估和对用户身份的持续积分评估，实现对访问主体的访问权限动态调整和访问身份认证动态调整。该技术模型具备分析、评估和决策三种能力，结构图如下：

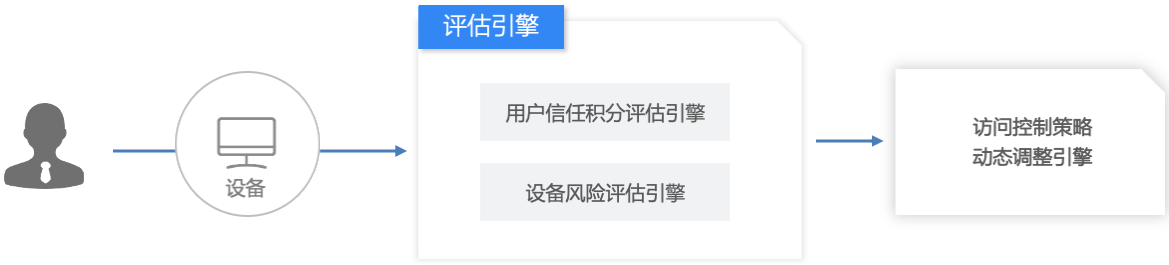

图 2 持续风险及信任积分评估模型

该技术模型包括两大核心智能算法模型：

- 用户信任积分评估模型：对所有访问主体进行例行的信任评估。每当用户发起业务请求时，根据访问主体的基础认证行为、环境因素、历史行为记录等多项属性情况进行综合评估，最终输出访问主体在当前一次访问过程的信任积分以及信任级别。在对用户的历史行为信任评估模型中，对于用户的异常时间、异常地点登录等场景的异常检测，采用了基于人工智能技术的长短期记忆网络算法（LSTM）、孤立森林算法以及离群因子检测算法；对于异常数据特征的提取，例如 mac 地址、ip 地址、用户 id、登录时间等等采用了基于人工智能技术的主成分分析算法（PCA）。

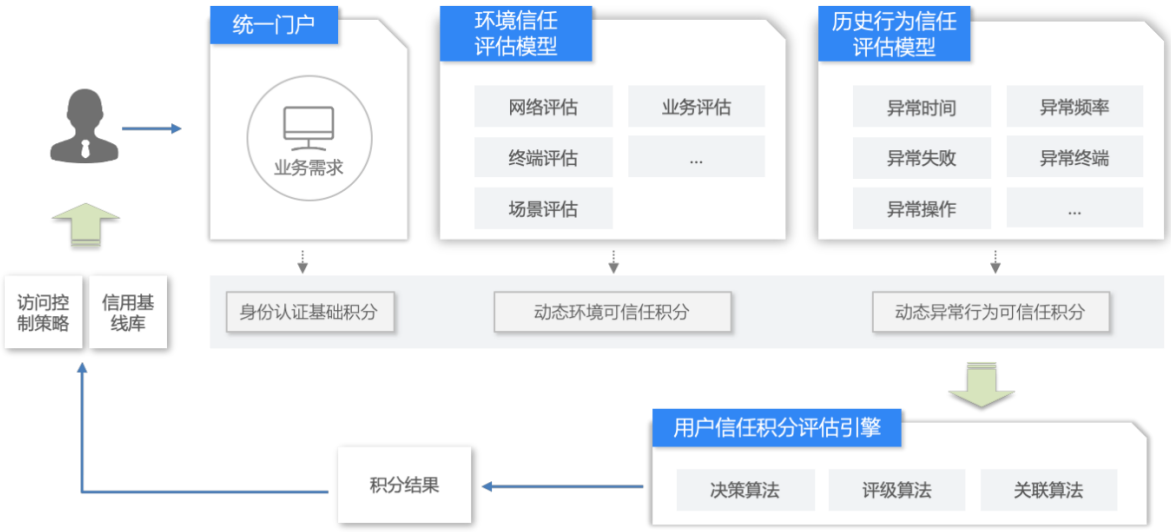

图 3 用户信任积分评估模型

- 设备风险评估模型：对所有访问客体进行周期性的信任评估。对设备的访问请求进行信任评估，根据设备的基础行为特征，结合动态调整模型，以及当前实体行为的异常风险级别，对访问客体进行综合风险评估，最终输出当前访问客体的安全级别以及对应的风险值。

图 4 设备风险评估模型

4. 访问控制策略动态调整引擎

访问控制策略动态调整引擎从持续风险及信任积分评估引擎中获得动态评估结果，例如访问主体的信任积分，系统风险值等，然后基于这些动态评估结果，向用户或设备提供动态的访问控制策略，访问控制策略可基于多种认证策略，包括认证级别升级策略、基于高风险操作的增强认证策略、特殊操作场景的多人联合认证策略、以及二次认证防护策略。

5. 多因素融合认证枢纽

多因素融合认证枢纽为企业用户或个人用户提供身份认证，支持主流多因素认证机制，包括知识因素、拥有因素、生物因素在内的各种认证方式，同时也可以转发用户认证请求到第三方认证代理。其中支持基于知识因素的认证机制，如密码、口令、问题等；支持基于拥有因素的认证机制，如令牌、手机号短信、邮箱账户、身份证号码；支持基于生物因素的认证机制，如指纹、人脸识别。支持第三方互联网认证服务的对接，例如微信、钉钉等。基于多种认证因素，可以按需设置多因素组合认证模式。

6. 可接入访问控制网关

可接入访问控制网关帮助企业在人员和资产设备之间搭建高效、可控的访问接入通道，为企业各类 IT 维护管理人员和第三方代维管理人员提供统一的接入维护入口，并对各类接入维护行为进行安全认证、鉴权、控制和审计等功能，可实现针对来源、人员、时段、行为、操作对象等多种细粒度访问控制，并对操作过程全程审计，形成审计记录、审计告警。

【应用效果】

该方案在效率和体验上，可以方便员工快速建立和接入工作环境，针对资产设备访问控制，人员身份验证等高频运维场景有显著的效率提升和安全体验：

资产设备一键可达：支持应用发布和帮助信息发布，对于用户办公、研发所需要的常用资源，用户可通过安全客户端快速打开所需资源。同时根据用户需要实时更新帮助内容，将用户所需常用操作指引、常见问题处理方式更新发布至安全客户端，快速触达用户。

快速身份验证：结合企业 SSO，实现快速、统一的身份验证，通过用户终端设备的一键登录，可以多次访问到业务管理平台、运维设备等。

【下一步工作建议】

零信任架构是系统工程，不可能一撮而就，首先需要从高层领导驱动，循序渐进的进行安全规划，方案选型保证开放化，选择新建业务优先推广。首先从用户认证安全入手，可以针对安全敏感的业务应用场景优化认证以及远程接入方案，并实现安全加固；其次基于零信任架构，针对办公应用（含移动办公）建立统一身份与访问控制，形成基础的零信任网络；然后建设统一身份、权限治理平台，形成具备工作流引擎和智能分析能力的身份治理平台；再然后将更多的业务接入零信任网络，逐步形成覆盖应用、设备、网络及流量的高级零信任网络；最后结合系统运营情况和上线数据，基于身份分析技术和工作流引擎推动零信任网络的持续完善和演化，从而推动企业安全重构，适应现代 IT 环境，助力企业数字化转型。

3.1.2 恶意代码分析

基于多机器学习模型的恶意代码智能分析检测

【场景描述】

近年来，网络空间面临严峻的安全威胁，所谓的“僵木蠕”（僵尸网络、木马、蠕虫）攻击是其中的典型代表。在这些攻击中，通常某种恶意代码会被作为攻击的载体，通过互联网、移动存储介质等传输至被攻击目标，运行后造成信息泄漏、数据损失、主机失陷等后果。例如，2017 年在全球范围内掀起一阵风暴的 WannaCry 勒索病毒，正是利用 Windows 系统的 SMB 服务相关漏洞进行传播，受害者主机上的图片、文档、音视频、压缩包、可执行程序等文件会被加密，如果受害者不支付赎金，被加密的文件将无法恢复。再比如，屡屡发生的 APT 攻击事件中，攻击者大多使用了鱼叉式钓鱼攻击的方法，诱导用户打开邮件中的附件或链接，从而触发后续的一系列攻击行为，造成受害者主机上的重要数据被窃取，或是系统被控制成为僵尸网络的一员等。在这些攻击中，作为攻击载体的恶意代码包括二进制 PE 文件（可执行文件）、PDF 文件、Office 文件、JavaScript 脚本等。现有的恶意代码分析检测技术主要基于特征指纹或规则库，对已知恶意代码的识别准确率高，但是通常无法识别新出现的未知恶意代码，而为了不断提升其检测能力就需要持续性投入大量人工来更新特征指纹或规则库。

针对以上问题，启明星辰发挥创新、研发优势，设计研发基于多种机器学习算法模型的恶意代码智能分析检测系统。该系统能够从样本中自动提取特征并进行分析检测，有效提升恶意代码分析检测的自动化水平，并对未知恶意代码有一定的检测识别能力，既减少了特征指纹或检测规则定义方面的人工投入，又可以通过持续使用新样本进行迭代训练的方式实现模型检测能力的自动提升。

【技术方案】

在恶意代码分析检测中引入机器学习技术符合有监督机器学习的一般工作流程框架（如下图）。根据不同恶意代码的类型，采用相应的预处理方法，再选择合适的人工定义与自动提取相结合的特征工程方法，所得的特征向量作为机器学习模型的输入，基于大量标注样本进行机器学习模型的训练，训练好的分类器应用于实际的恶意代码检测与分类中。其中，使用的机器学习模型可以由多个不同的子模型所组成，通过集成学习的方法对多个子模型结果进行再学习以输出最终结果。

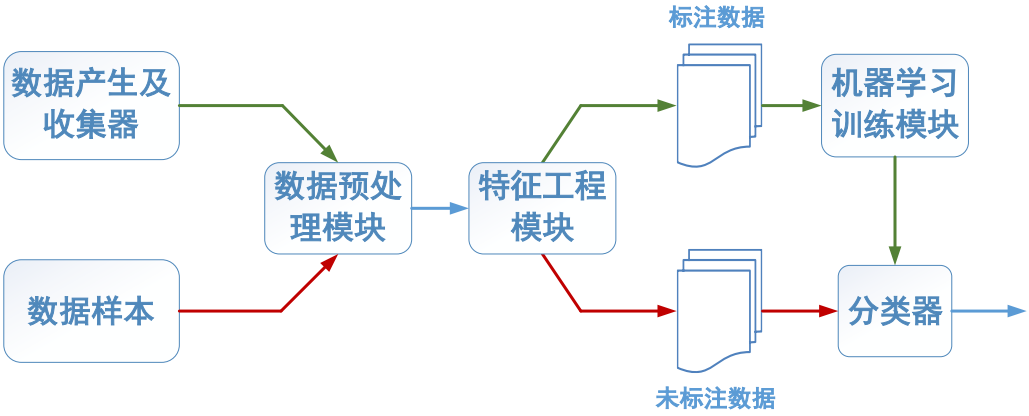

图 1　一般有监督机器学习的工作流程框架

下面以 PE 二进制恶意代码类型为例，给出一个具体的技术方案描述。

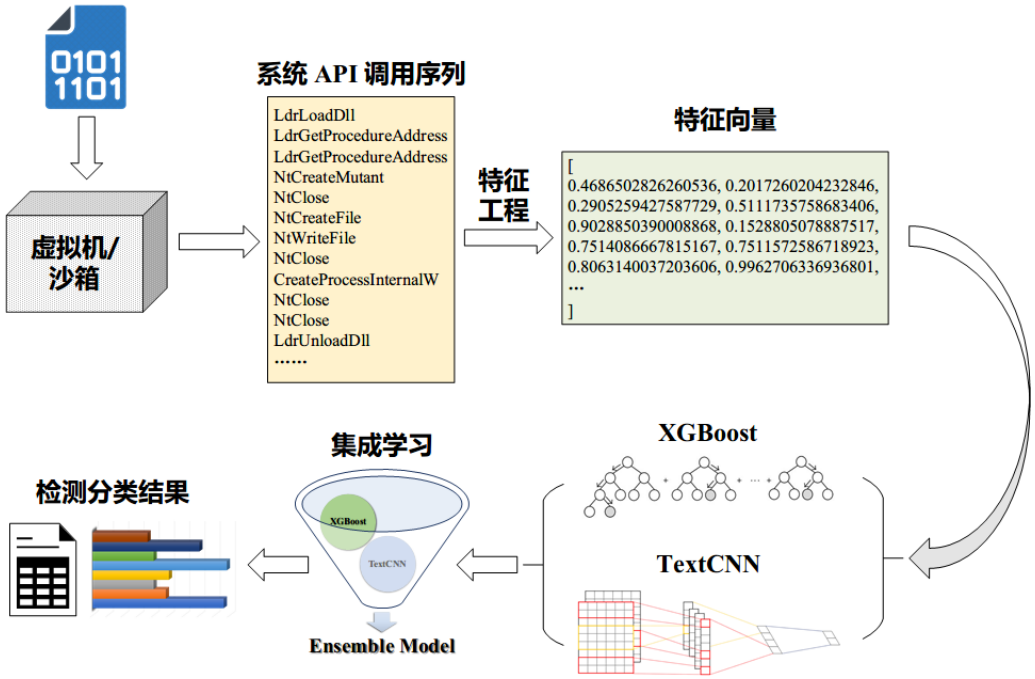

图 2　基于机器学习的 PE 二进制恶意代码分析检测方案

首先，将待检测的 PE 二进制格式可执行代码在虚拟机或沙箱中运行，监控并记录所产生的系统 API 调用序列；然后，对获得的 API 序列进行特征工程，也就是使用 N 元模型、Word2vec 模型等方法将 API 序列转换为数值向量（称为特征向量）；接下来将特征向量作为机器学习模型的输入，这里采用了两种不同的模型，分别是基于提升树的传统机器学习分类器 XGBoost 和基于卷积神经网络的深度学习分类器 TextCNN，均需要使用大量标注数据进行离线的训练；获得两个分类器的输出结果后，再应用集成学习的方法对分类结果进行再选择，从而得到最终的分析检测结果，即输入的代码是否为恶意代码及其具体类型。

【应用效果】

经过实验环境以及大量公开数据集的测试评估，基于多机器学习的恶意代码智能分析检测系统可以有效提升恶意代码的分析检测自动化水平及检测准确率。对于几种常见类型的恶意代码，其检测准确率如下表。

恶意代码类型	检测准确率
PDF/Office	98.8%
JavaScript	99.0%
PHP WebShell	98.4%
二进制 PE	98.7%

此外，该系统对未知恶意代码具有一定的检测能力，具体的检测成功率与未知恶意代码和已知恶意代码的相似度有关，在一次实验评估中对训练数据集中未出现种类的恶意代码的平均检出率能够达到约 70%。

本方案中所采用的相关技术已经提交发明专利申请 1 件，待发表论文 1 篇。

【下一步工作建议】

持续积累最新的、不同类型的恶意代码样本，对算法模型进行优化，并构建模型迭代更新的机制，根据模型的实际表现确定模型重新训练的频率。另外，需要研究如何在保证准确率的前提下简化机器学习算法模型，降低计算复杂度，以更好地适用于大数据环境，提供满足离线分析、在线检测等不同场景需求的分析检测机制。

一种层次化的机器学习引擎恶意代码检测方案

【场景描述】

当今网络环境下，新型恶意威胁层出不穷，这对恶意代码检测提出了极大的挑战。传统的威胁检测引擎基于规则匹配进行威胁检测，其能力依赖于规则的时效性与通用性。对于未知威胁，传统引擎在规则没有及时更新或是通用性不佳的情况下会存在漏报的风险。

亚信安全提出了一种层次化的机器学习引擎恶意代码检测方案，该方案通过机器学习技术关联威胁信息和深入分析文件，通过提取文件的常用特征来检测新出现的未知安全风险。该方案还会对未知或不常见的进程进行行为分析，以确定是否有新出现或未知的风险正在尝试感染网络。同时，该方案引入了一种层次化的识别体系，将亚信安全已有的恶意代码检测引擎与预测型机器学习引擎有机结合，通过前置引擎分析与后置过滤，可以有效的提高恶意代码检测率，减小误报率，从而更好的实现终端环境的全方位保护。

【技术方案】

一、技术方案概述

本方案整体架构如下图所示：

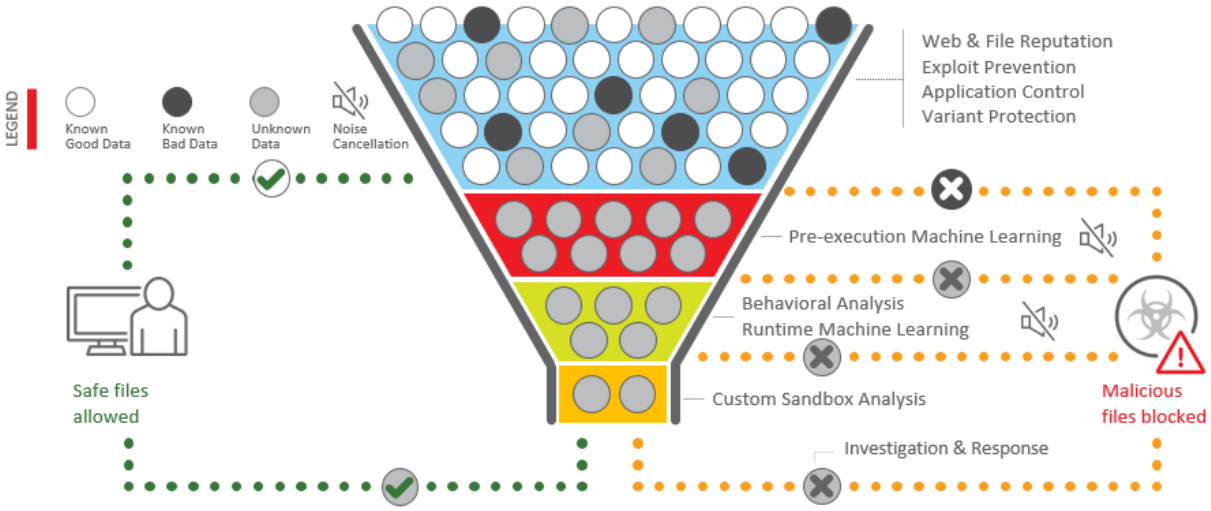

图 1 方案整体架构

如上图所示，对于恶意文件检测，该方案包含以下四个层次：

1. 待检测文件先经过已有的传统引擎进行扫描，传统引擎会将已知正常文件与已知恶意文件滤除，只留下未知文件进入下一层处理。

2. 未知文件由机器学习引擎进行识别，在未执行的情况下，机器学习引擎会提取未知文件的静态文件特征，并结合相应的静态模型与识别算法判断未知文件是否是恶意文件。如果识别正常，则进入下一层进行判断。

3. 当未知文件被用户运行起来后，机器学习引擎则会提取未知文件相关进程的行为特征，并对其进行实时分析。通过结合动态模型与进程实时动作，可以在未知进程有损害用户安全的动作出现前及时阻断未知进程，隔离未知文件。

4. 用户同样可以将未知文件放入自定义的沙箱中，在可控环境中运行未知文件。通过对未知进程的运行时行为与结果的全面记录与分析，可以最终判定该文件是否是恶意文件。

二、机器学习引擎工作流程：

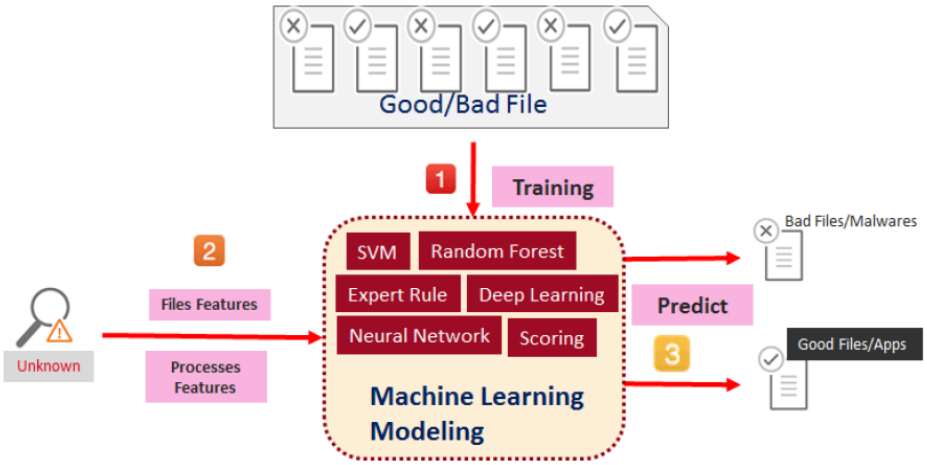

图 2 机器学习引擎工作流程

1. 训练流程：引擎收集及整合已知的正常与恶意文件，将其作为训练样本，采用已有模型进行训练，生成相应的 Pattern。

2. 特征提取：针对未知文件，引擎对文件进行特征提取并进行进一步处理。

3. 模型预测：结合训练生成 Pattern 与未知文件特征，通过采用相应机器学习算法，对未知文件进行分类。

二、方案优点

该方案具有以下的优点：

1. 通过结合传统引擎与新型机器学习引擎，可以有效的在检测已知威胁的同时具有对未知恶意威胁的识别能力。通过前置传统引擎，分层处理，可以高效的对正常与恶意文件进行过滤，从而使得机器学习引擎能够更专注于未知威胁的识别。

2. 噪声消除引擎可以通过给定文件的成熟度与流行度综合判断文件是否是恶意文件，通过结合噪声消除引擎与机器学习引擎，可以有效的降低误报率，从而减小文件误报带来的影响。

3. 机器学习引擎从静态与动态两个方面对未知文件进行识别，可以有效的避免恶意文件静态变形逃逸检测，因此可以提供更全面的保护。

【应用效果】

该方案作为亚信终端防护产品 OSCE 的核心组件之一，有效的保护了终端用户的系统安全。在 2017 年勒索软件 Wannacry 爆发事件中，该方案在其它传统检测引擎失效的情况下成功的检测出了 Wannacry 勒索病毒并在第一时间阻断了其对用户文件的加密，避免了未知恶意软件对客户的系统造成进一步的伤害。

【下一步工作建议】

对于机器学习引擎，可以进一步优化算法模型，同时持续更新训练样本集，使其具有更高的识别率。另一方面，可以进行横向扩展，使得该引擎可以支持更多类型与场景的恶意文件识别。最后，该层次化的体系可以根据实际用户场景进一步细化，通过加入更多的中间层与专用引擎，增加其在各种安全场景下的适用性。

基于机器学习的未知恶意代码检测

【场景描述】

传统的 APT 防护技术专注于从企业客户自身流量和数据中通过沙箱或关联分析等手段发现威胁。由于企业网络防护系统缺少相关 APT 尝试分析经验，而且黑客躲避安全检测、绕过安全防御的水平也在不断的提高，传统检测手段会经常性的出现误报和漏报现象。

恶意文件、病毒木马是黑客入侵主要采用的攻击实施方式，传统的恶意文件检测大多基于签名特征库匹配机制。在实际入侵过程中，攻击者会采用乱序、加壳、动态生成的方法消除恶意文件中可供检测的特征，甚至利用 0Day 漏洞对应的恶意代码，直接绕过传统文件检测手段。

为解决此问题，奇安信根据积累的海量恶意文件供检测引擎进行机器学习，积累远超特征码能力的恶意文件检测模型，大幅提升对 0Day 漏洞恶意文件、免杀恶意文件的检测能力。

【技术方案】

基于人工智能的杀毒引擎，依靠海量数据挖掘、引入机器智能学习算法，能够有效准确识别未知恶意软件，能够根据已知的正常软件和恶意软件的大量样本，通过数据挖掘找出两类软件最具有区分度的特征，建立机器学习模型，使用机器学习算法，得到恶意软件的识别模型。通过获得的模型对未知程序进行分析判断，即可获得软件的恶意概率，从而在可控的误报率之下尽可能多的发现恶意程序。

机器学习引擎的学习流程如下图所示：

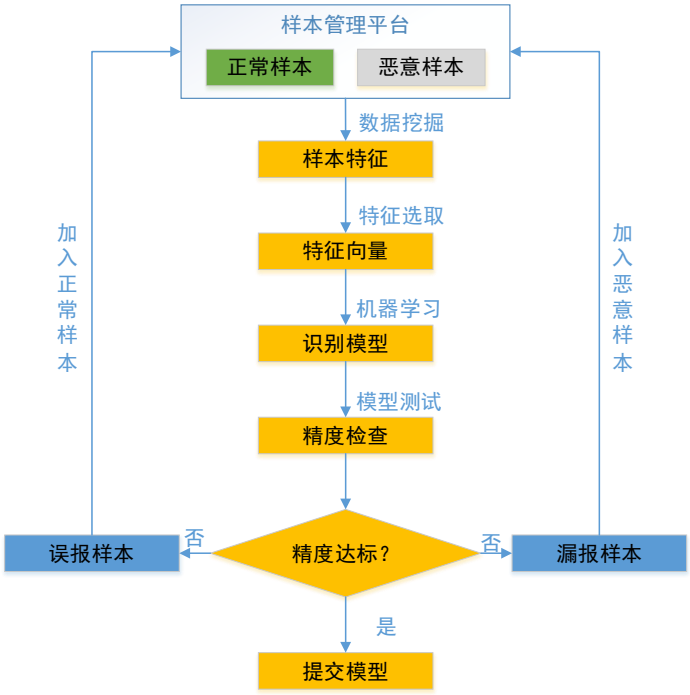

图 1 机器学习引擎流程示意图

通过对训练样本的数据挖掘，例如导入 API 函数、PE 头部信息、代码反汇编信息等等进行海量数据挖掘，找到海量 PE 文件特征。应用特征选取算法，选取最有效的特征，建立特征模型。

利用特征模型对训练样本数据进行数据特征化变换，生成对应的特征向量，利用成熟的机器学习算法（例如 SVM)，对样本进行训练，得到恶意程序识别问题的识别模型。

对生成的模型进行测试，如果精度达到要求，则终止。否则对误判样本进行分析，调整样本的分类属性，再次迭代。

该引擎的运行环境如下图所示：

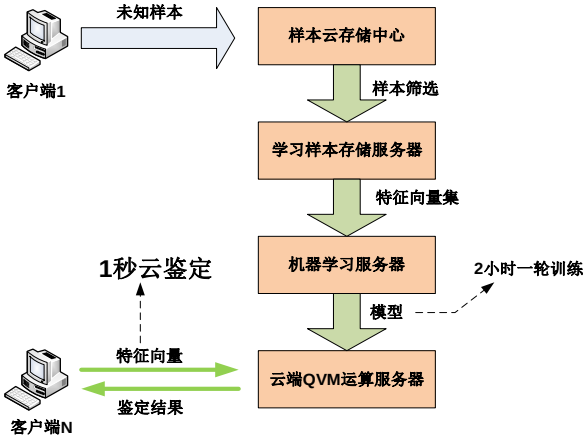

图 2 机器学习引擎运行环境示意图

本技术首要成功要素是低误报率。根据前期研究的结果，一个合适的机器学习算法的选择对误报控制是相当重要的，采用 SVM 作为基本学习算法，并设计了快速的参数选择，和快速训练方法。

机器学习算法在人工少量干预样本（添加、删除、修改黑白属性)指导的情况下，系统能够实现自我学习，自我进化。目前该引擎学习一轮的时间仅为 2 小时。

机器学习有效的解决了大部分未知恶意程序的发现问题。由于传统杀毒技术严重依赖于样本获得能力和病毒分析师的能力，基本只能处理已知问题，不能对可能发生的问题进行防范，具有严重的滞后性和局限性。本技术对海量样本进行挖掘，能够找到恶意软件的内在规律，能对未来相当长时期的恶意软件技术做出前瞻性预测，实现不更新即可识别大量新型恶意软件。

机器学习使得对样本分析人员的要求相对较低，仅仅需要分析员能够区分文件是否恶意，而不需要人工分析恶意软件实现方法和识别方法，降低了人员参与门槛，大大节约了人力成本。

【应用效果】

本案例可应用于国内外大量重要客户的网络边界和重要系统关键网络节点，包括运营商、金融、医疗、教育、能源、电力、国家机关等。这些客户在所面对的黑客攻击强度大、攻击水平高，传统安全检测手段无法有效应对 0Day 漏洞恶意样本、免杀恶意样本。而黑客在生成攻击样本中采用了自动化构建免杀恶意文件的方式，使得每时每刻都会有大量全新的难以检测的样本产生。借助人工智能技术可以大幅提升检测的能力，有效减少对专业安全分析人员的依赖，减少低效重复的人力劳动。

【下一步工作建议】

具备人工智能特性的 APT 检测技术是新一代安全威胁检测的必备方案，能够将将企业的防护水平从基于已知规则、依赖人工分析的水平提升到自动化、智能化高度。建议优先在关键信息基础设施运营单位开展相应的技术推广工作。

3.1.3 恶意域名检测

引入数据增强技术的智能 DGA 域名检测

【场景描述】

在早期的僵尸网络中，控制者通常把 C&C 服务器的域名或者 IP 地址硬编码到恶意程序中，僵尸主机通过相关信息定时访问 C&C 主机获取命令。但同时安全人员也能够通过逆向恶意程序，得到 C&C 服务器的域名或者 IP，利用这些信息定位 C&C 主机，从而阻断僵尸网络。黑客为了保护 C&C 主机，使用 DGA 算法产生大量备选域名，使得无法通过静态规则来进行检测，因此基于机器学习的 DGA 域名检测技术被提出。但是现有的基于机器学习的 DGA 域名检测系统仍存在两方面的不足：一是用于训练模型的数据集中 DGA 域名往往远多于正常域名，导致系统在实际应用中的误报率较高；二是部分 DGA 算法生成的是与正常域名更为相似的低随机性域名，导致系统的漏报率升高。

针对以上问题，启明星辰设计研发了一种新的智能 DGA 域名检测系统，通过引入数据增强技术解决了正常域名数据量较少的问题，同时采用新型深度学习模型——基于注意力机制的双向 GRU（门控循环单元）神经网络，取得更高的检测准确率和更低的误报率。

【技术方案】

整体方案包含如下四个步骤：

- 收集正常域名和 DGA 域名数据，然后对数据进行预处理；
- 使用数据增强技术，增加正常域名的数据量；
- 训练基于注意力机制的双向 GRU 神经网络模型；
- 模型线上部署，实现实时预测。

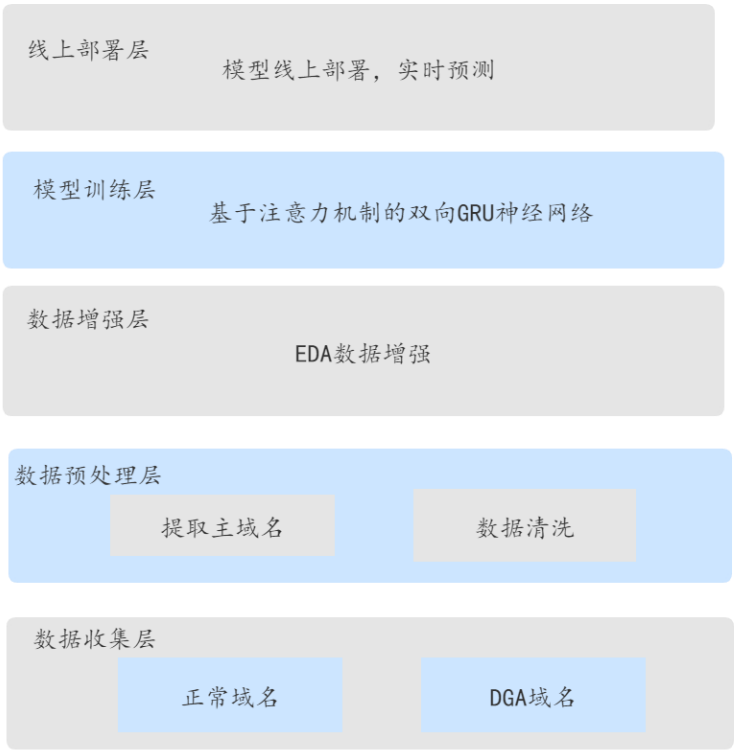

图 1 引入数据增强技术的智能 DGA 域名检测系统结构图

收集数据，分为正常域名和 DGA 域名的收集。其中正常域名数据主要来源于 Alexa、Umbrella 等网站排名前 100 万的域名，并结合威胁情报将其中的恶意域名去除；DGA 域名数据主要来自于威胁情报，少量由开源的 DGA 算法自动生成。

数据预处理，包括主域名提取、数据清洗等操作。其中数据清洗主要是将同属于正常域名和 DGA 域名的数据删除。

应用数据增强技术，增加正常域名的数据量。具体来说，采用自然语言处理领域常用的 EDA（Easy Data Augmentation）技术，通过选取已有的正常域名并对其进行少量的随机字符添加、删除或替换操作以制造

出大量新的域名作为补充的正常域名样本，使得训练数据集中正常域名与 DGA 域名的数量比例减小，从而缓解数据不均衡的问题。

采用基于注意力机制的双向 GRU 神经网络模型（结构如下图），相比于传统的机器学习模型，这一模型在对一些 DGA 算法产生的低随机性域名检测上有更好的表现。同时，为了进一步提高模型的泛化能力，在模型训练构建中应用多种技术，包括在训练前使用 Word2vec 得到每个字母对应的词向量作为嵌入层的输入、加入 Dropout 防止过拟合等。

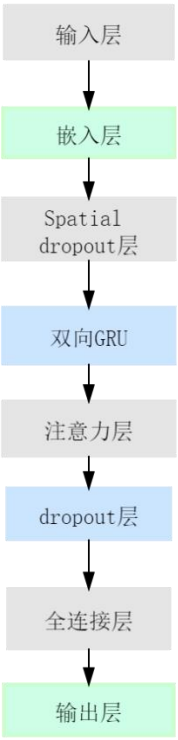

图 2 基于注意力机制的双向 GRU 神经网络模型结构图

将训练所得的模型进行线上部署，是指将模型预加载并持久化到内存中（后台服务），之后通过网络通信方式（RESTful API）将待检测数据发送至后台服务，模型对接收到的数据进行预测并返回结果。

【应用效果】

在实验测试环境下，使用约 100 万条数据对新模型进行测试，相比于原有的 DGA 检测系统，误报率从 4.2%降低到 0.5%，漏报率从 3.2%降低到 0.3%，从而大大降低了误报率和漏报率。

【下一步工作建议】

本方案中采用的神经网络属于深度学习模型，虽然检测效果更好，但其所消耗的计算资源也会急剧增加。因此，为了兼顾效果和性能，在未来将尝试使用模型蒸馏和模型量化等技术，对已有模型进行压缩，从而降低对计算资源的需求。

兼容拼音域名的多层自适应 DGA 域名检测方法

【场景描述】

在攻击活动中，当感染了恶意代码的宿主机要与 C&C 服务器联络，以获得进一步攻击活动的指示时，如果将 IP、domain 通过硬编码方式写入代码中，很容易被有经验的安全人员或数据威胁感知发现。在此背景下，攻击者会利用 DGA（domain generation algorithms）域名生成算法，使用写入代码的随机算法和约定好的随机种子生成一系列的域名，攻击者一般只注册域名列表中的一小部分，被感染的宿主会挨个尝试请求这些域名，直到可以解析出 IP，连接 C&C 服务器。现有的域名过滤方案，通常可能存在如下不足之处：

- 若使用传统的黑名单方法，由于 DGA 域名生成的随机性强、数量众多，很难进行有效的防控。
- 基于 DGA 域名的字符随机性设计的机器学习检测模型，对现在比较流行的从单词列表中生成的 DGA 域名，难以进行有效的识别。
- 基于英文域名样本训练的模型，在中文场景下存在一定的误报概率。

针对以上问题：本方案提供的检测算法是针对 DGA 域名的多样性、多语言性的一套静态特征的组合算法，其中 N-gram 模型针对随机字符型 DGA 域名，LSTM 模型针对单词组合型 DGA 域名，拼音特征和拼音加强数据集是针对中文语境下的加强，最后将多模型组合起来综合评估。

【技术方案】

一、技术方案概述

本方案整体系统架构如下图所示：

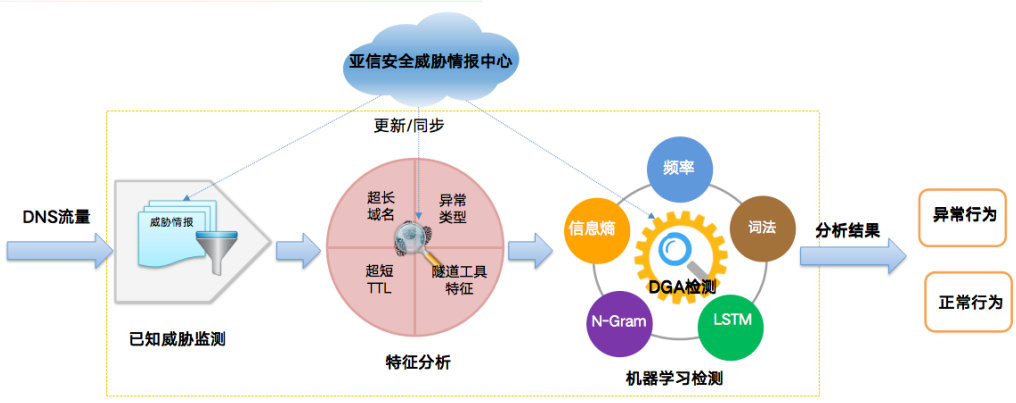

图 1 系统架构图

二、整体方案步骤

整体方案包含如下步骤：

- 根据已有威胁情报实现已知威胁实时监测；通过域名特征分析以及基于机器学习的多种分析算法，快速检测出企业内部的疑似恶意域名威胁。
- 定期从亚信安全威胁情报中心下载更新威胁情报、域名黑白名单和机器学习预测模型。

其中，DGA 域名检测系统的技术模型如下图所示：

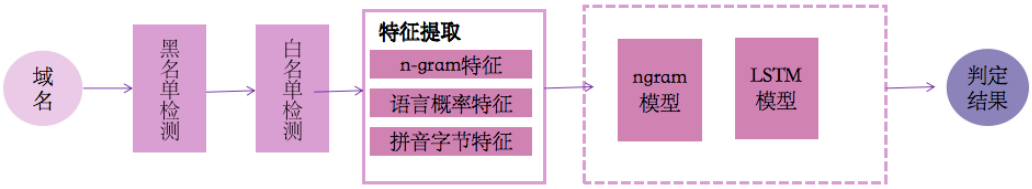

图 2 DGA 域名检测系统的技术模型图

三、技术优势

该技术模型包括如下优势：

- 特征提取：

本方案针对不同的 DGA 家族生成机制，设计了不同的特征提取方法：包括长度、信息熵，针对随机字符型 DGA 域名提取的 n-gram 特征，针对单词组合型 DGA 域名提取的语言模型概率特征，还有针对中文语境下的域名，提取的拼音字节特征。

- 中文域名特征：

在测试中，我们发现原有的方法和特征对于中文语境下的域名，特别是由拼音拼成的域名，存在一定的误报，原因可能在于基于英文域名、英文环境的算法在中文语境下会有一定的泛化问题。为此，我们在本方案中，特别设计了针对中文拼音域名的拼音字节特征，并增强了拼音型域名的正常数据集。这些措施，根据测试，有效地缓解了拼音型域名的误报。

- 分层、综合判定模型：

本方案针对 DGA 的多样性、多语言性，提出了一套静态特征的组合算法，其中 N-gram 模型针对随机字符型 DGA 域名，LSTM 模型针对单词组合型 DGA 域名，拼音特征和拼音加强数据集是针对中文语境下的加强，再综合相应的特征指标和模型判定，得到最终分析结果。该方案对多语言，变化多样的 DGA 域名有较强的覆盖能力，相比于基础模型，对单词组合型 DGA 域名检出的准确率有较大提升，而误报率，特别是针对中文语境域名的误报率，也有很大改善。同时该方案也是一个根据待检测域名的相应特征指标，动态选择所需模型（是否需要引入消耗资源较多的深度学习模型）的分层次解决方案。这样的方案，经过测试，可以有效地提升性能，降低资源消耗。

- 线上判定与线下训练相结合：

该技术模型包含线下训练和线上过滤判定两个流程：线下训练流程基于自有数据集进行综合训练，并会定期收集线上判定的反馈数据，重新训练模型，并对新模型进行评估，若新模型的表现超越旧模型，则进行替换。线上过滤判定流程则依据 N-gram，LSTM 模型等对可疑域名进行综合判定。

【应用效果】

本方案部署在亚信安全威胁情报中心 Maldium 引擎，性能可达到数千 QPS，日均检测疑似 DGA 域名约占总域名数的 3%。

【下一步工作建议】

DGA 域名检测是恶意域名的一种重要的可疑指标，更进一步，可以结合域名注册特征，域名关联样本特征等，对域名的恶意性做更全面综合的判定。

针对 DGA 隐蔽域名的人工智能发现机制

【场景描述】

DNS 服务是互联网中最基本、最常用的网络协议，由于其基本性和通用性，也容易被安全攻击手段利用。主要被用于网络攻击的控制和传输阶段。

在控制层面，攻击者会控制僵尸网络发起攻击，即攻击者需要告知僵尸主机命令，而此命令的下发通过命令控制信道（C&C Channel）实现。在早期中心结构的僵尸网络中，僵尸主机通常采用轮询的方法访问静态硬编码 C&C 域名或 IP 来访问命令控制服务器，获取攻击者命令，由于硬编码的域名或 IP 固定且数量有限，安全防护人员通过逆向分析木马文件的样本即可掌握这部分域名和 IP 地址，利用威胁情报手段，可以让这部分域名和 IP 地址在全球范围内被公开并屏蔽。如果指定的 C&C 域名、IP 地址不可访问的，控制者就失去了对整个僵尸网络的控制能力，因此 C&C 是僵尸网络构建的核心，也是攻防双方博弈的关键点。

为了克服可能被批量屏蔽的情况，攻击者使用 Domain Flux 协议来对抗防御人员的域名屏蔽，僵尸主机访问的 C&C 域名不再是静态硬编码，而是根据一定算法动态生成的、变化的域名，该域名生成算法称之为 DGA（Domain Generation Algorithm），算法的输入称为 Seeds，涵盖日期、社交网络搜索热词、随机数或字典。对安全防护人员来说，DGA 识别有着重要的安全应用价值。

【技术方案】

奇安信在 APT 检测系统中实现了多种基于行为、数据分析的异常检测 DGA 域名检测。

基于流量的方法，主要针对于 DGA 算法，通常表现为生成海量域名，但其中只有少部分可被正常访问的特点，进行构建。如通过分析反复出现的“动态”的应答为 NXdomain 的 DNS 请求，实现 DGA 域名的识别。以及在给定时间内某域名被多个感染主机查询。虽然基于网络流分析或深度报文检测的僵尸网络检测系统可以在本地网络内检测到感染主机，但是它们不能针对大型 ISP 环境下的大规模流量进行很好的扩展。

相比于其上传统方法，机器学习可更高效的利用专家经验，同时减少甚至无需依赖网络流量数据等三方数据，来完成 DGA 域名的检测任务。如基于分类、回归等算法，均可高效准确的判别一个域名是不是 DGA。

此上算法均属于监督学习，其算法基本流程为：

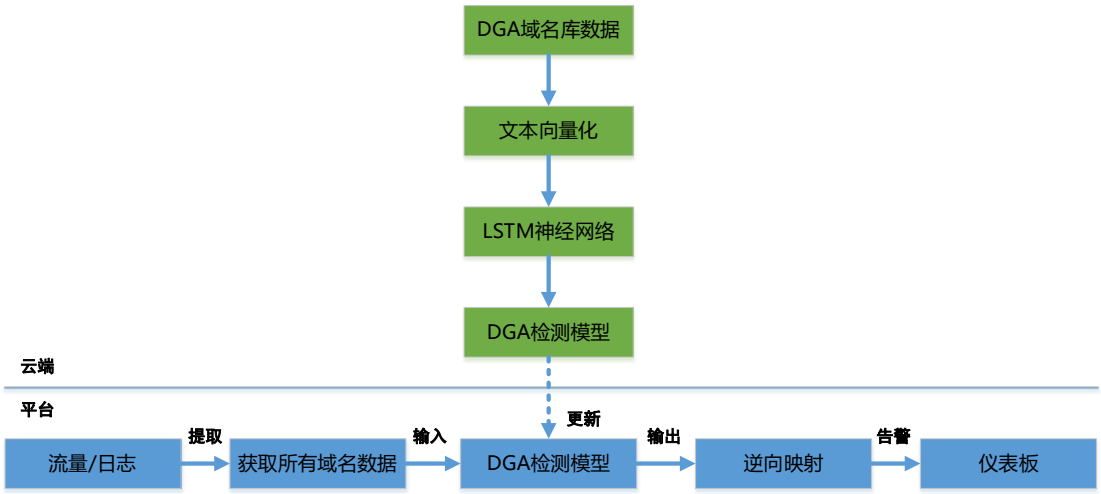

图 3 云端与本地结合的 DGA 域名检测模型训练方式

1. 在云端进行基于 DGA 域名库数据集的训练；
2. 把人类的经验表示为特征把数据集转换成特征向量；
3. 利用这些数据集和他们的特征向量训练合适的算法模型；
4. 评价算法效果，比如精度、召回率等等，并交叉检验分类效果；
5. 将生成的 DGA 检测模型从云端下发至部署在各处的 APT 检测设备中；
6. APT 检测设备通过流量采集、日志分析的方式获得网内查询过的域名数据；
7. 将域名数据输入 DGA 检测模型进行逆向映射分析；
8. 如符合 DGA 检测模型的告警级别，则判定该域为 DGA 域名并告警。

在此 DGA 识别算法中，训练及测试数据仅有网站域名无需其他数据，正负案例数据分别采用 alexa 排名前 100 万域名与奇安信积累的 DGA 域名。

在算法特征处理阶段主要模拟了网络安全专家人工判断 DGA 时的“直觉”进行设计，如域名的随机性、域名的连续分散度、语言学相关知识等。依靠以上高维特征，将域名映射至专家经验空间，为 DGA 域名的高效准确识别打下坚实的基础。

在算法判别阶段，基于随机森林、深度学习、KNN 等算法，充分利用输入特征，进行模型融合后，最终实现 DGA 域名的高效准确识别。

表 1 各种人工智能算法在 DGA 域名中的检测效果

计算方法	检测效果
随机森林	经典随机森林分类结果按分类树投票多少形成的分数而定。在随机森林中单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类。
马尔科夫链	在 DGA 检测中，马尔科夫链主要应用于域名在 n-gram 切分后的 n-gram 常见性计算上，寻找发现可疑域名。
深度学习	深度学习即深度神经网络，在此主要应用的方法为 RNN，由于其特殊的算法特点，在众多自然语言处理中取得了巨大成功以及广泛应用。其算法特点也与 DGA 判别任务极为契合。
k近邻分类(KNN)算法	基于 KNN 算法，通过调整各特征维度权重，在特征空间区将正常域名与 DGA 区分下，可实现对于新域名的有效划分，完成 DGA 识别。

【应用效果】

本项目应用于国内外大量重要客户的网络边界和重要系统关键网络节点，包括运营商、金融、医疗、教育、能源、电力、国家机关等。黑客在针对这些客户成功入侵后，会采用隐蔽通道进行远程控制和数据传输，DGA 域名由于具有无法提前获取、与正常域名难与区分的特点，被黑客广泛用于入侵后的远程控制和数据传输阶段。云端与本地结合的人工智能技术能够有效筛选出 DGA 域名，通过对外请求的 DNS 查询快速判断内网主机是否失陷。

【下一步工作建议】

具备人工智能特性的 APT 检测技术是新一代安全威胁检测的必备方案，能够将将企业的防护水平从基于已知规则、依赖人工分析的水平提升到自动化、智能化高度。建议优先在关键信息基础设施运营单位开展相应的技术推广工作。

3.1.4 恶意流量识别

Webshell 通信流量智能检测与规则自动化提取

【场景描述】

Webshell 是以 PHP、JSP、ASP 等网页文件形式存在的一种命令执行环境，也被称为网络后门。通常攻击者利用网站漏洞将 Webshell 后门文件传输至网站服务器，与正常网页文件混在一起，之后就可以通过浏览器访问的方式实现对网站服务器的控制及数据的窃取。传统的基于网络流量的 WebShell 检测方法主要是通过对报文中出现的一些特征字符串进行匹配，攻击者可以通过代码修改或使用代码混淆等方法实现绕过检测。另一方面，Webshell 的种类繁多且不断出现新型 Webshell，对其进行人工分析以提取特征费时费力。

对于以上问题，启明星辰发挥创新、研发优势，设计研发了一种基于决策树算法的智能检测技术及规则自动化提取方法，能够自动分析 Webshell 样本并采集其通信流量，基于从 HTTP 报文提取的特征构建决策树模型，实现精准的 Webshell 检测，并可通过模型自动提取检测规则，应用于传统检测引擎中。

【技术方案】

本方案的主要目标是利用人工智能技术提高 Webshell 样本分析的效率，实现对样本通信流量的自动化采集与特征提取，构建精准的检测模型并支持自动化规则抽取。技术方案的具体流程示意图如下。

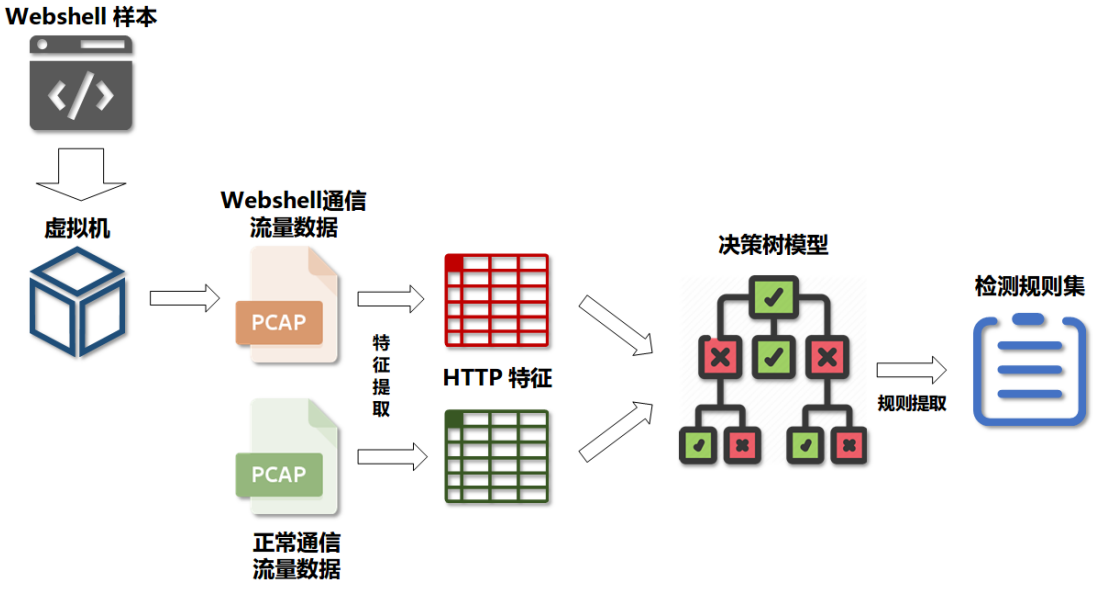

图 1Webshell 智能检测与规则自动化提取技术方案示意图

首先，将获取到的 Webshell 样本部署在虚拟机中，然后通过自动化的程序进行操作与其交互，采集监听到的通信流量；对这些 Webshell 的通信流量数据与其他正常网络通信流量数据进行特征提取，得到 HTTP

协议报文的数值特征集；基于提取出的特征应用决策树模型进行训练，得到可用于检测 Webshell 的分类器，而且可以基于决策树导出检测规则，进行归并与筛选后可直接应用于基于规则的检测引擎。

本技术方案已经提交 2 件发明专利申请。

【应用效果】

在实验测试中，针对约 100 种不同类型的 Webshell 样本（包括的网页版 Webshell 及 Webshell 管理工具）进行自动分析处理，采集了总计 100 多万条网络通信的 HTTP 报文，构建的决策树模型在十折交叉验证中可达到平均 99.9%的检测准确率。对从决策树模型中自动提取的检测规则进行归并与筛选，得到在验证数据集上无误报的检测规则 100 余条，可选用于补充传统检测引擎的规则库。

【下一步工作建议】

持续收集新的 Webshell 样本，采集更丰富的通信流量，以构建更细粒度的检测分类模型，实现对具体 Webshell 类型的准确识别，为攻击同源性分析提供情报信息。另外，可将本方案中的自动分析恶意样本、构建模型及提取规则的方法扩展，应用于其他网络攻击、恶意流量检测等问题中，实现自动化分析学习新的恶意流量样本并生成检测规则，为安全研究人员提供参考，以减轻其工作量，提高样本分析效率。

3.1.5 智能安全运维

基于人工智能的自动化响应与处置系统

【场景描述】

传统 SIEM 存在告警过多，客户关注的、有效的告警被淹没，无法对安全运维及研判人员起到有效的帮助等问题；同时企业缺乏专业的安全攻防、分析、处置人员，且员工在安全分析研判上的经验难固化；除此之外，传统运维方式存在响应、处置时间过长，效率较低，缺少规范化的响应处置流程和可量化的指标等缺陷。

针对上述挑战，绿盟科技将人工智能技术与专家经验进行有机整合应用于，推出智能安全运维系统实现从安全分析到响应处置全流程闭环，人工智能技术的引入事件分析研判、可视化编排等技术之中，通过与专家经验进行有机整合，推出智能安全运维系统。系统可显著提升威判的准确率，有效降低威胁事件的误报、漏报，加快安全事件响应处置速度，在保障业务安全的基础之上，提高了运营效能，实现从安全分析到响应处置全流程闭环。

【技术方案】

一、系统架构

绿盟科技智能安全运维系统，系统技术架构如下：

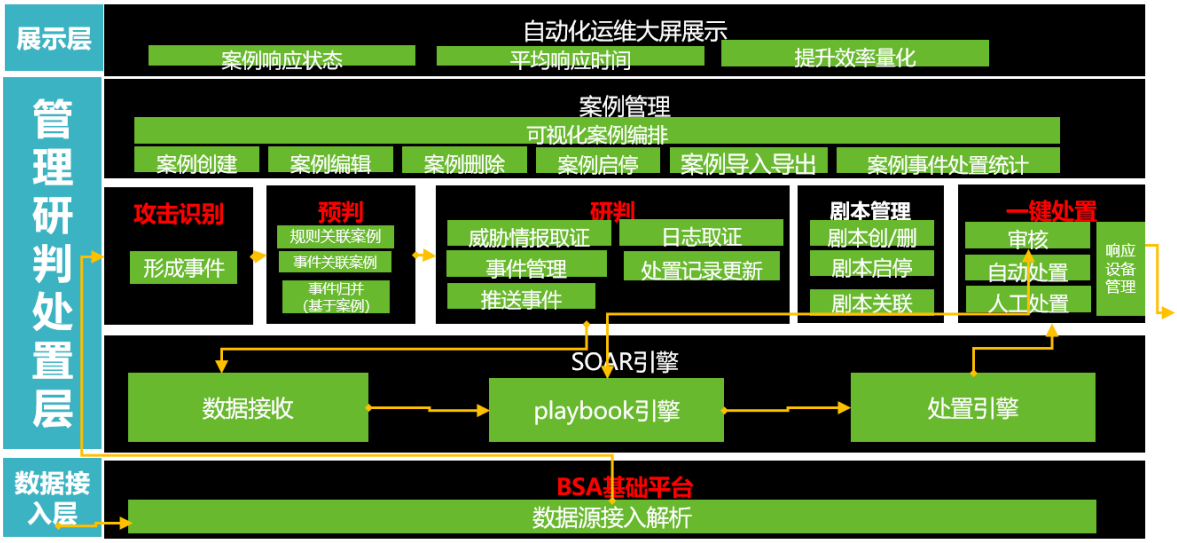

图 1 智能安全运维系统架构

- 数据接入层：主要实现分析数据的接入及解析，需复用基础平台。
- 管理研判处置层：该层是 SOAR 模块的核心，实现安全编排及响应处置决策及基础功能管理，涉及的关键组件如下：
  - 研判：在 SOAR 系统中，支持网络层面的取证，包括威胁情报取证、流量取证、主机取证；同时根据预判引擎推送的事件与案例关系信息，需要自动化处置的事件推送至 SOAR 引擎，同时会记录 soar 引擎处置的结果，在运维事件中做集中展示。
  - 剧本管理：实现对处置剧本的管理，包括剧本可视化编排、剧本创建、剧本删除、剧本编辑、剧本启停、剧本关联等。
  - 一键处置：为 SOAR 提供响应能力，支持接收 SOAR 下发的自动处置动作，通过插件定义并管理响应设备。
  - SOAR 引擎：基于研判输出结果，发送需要处置的安全事件，将事件解析后，并选择相应的 Playbook 剧本，生成 action 下发到一键处置模块，实现设备联动处置闭环。
  - 案例管理：对涉及的案例进行集中管理，包括创建、删除、编辑、启停、导入、导出等操作
- 展示层：通过自动化运维大屏对 SOAR 一些指标进行全局展示，将不同运维指标进行量化。

绿盟科技智能安全运维系统利用人工智能技术提供自动化编排与响应逻辑功能，具体流程如下：

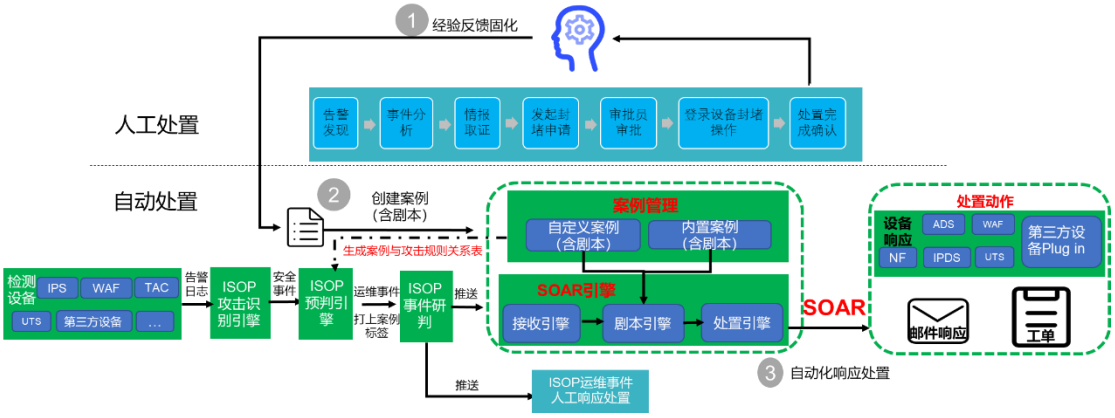

图 2 自动化编排与响应流程

1. 通过可视化编排技术将人、安全技术、流程进行深度融合，通过 Playbook 剧本串并联构建安全事件处置的工作流，自动化触发不同安全设备执行响应动作；

2. 基于对安全事件上下文更全面、端到端的理解，有助于将复杂的事件响应过程和任务转换为一致的、可重复的、可度量的和有效的工作流，变被动应急响应为自动化持续响应。

3. 案例管理功能将安全运维工程师经验固化，形成了攻击识别规则、运维事件与自动化响应处置动作之间的关系，一旦攻击事件与 SOAR 中开启的案例得到了匹配，该事件可在无需人工干预的情况下，通过剧本完成自动化闭环处置。

自动化编排与响应逻辑功能人工智能的引入将高级工程师从日常运维中释放出来，既可以缩短响应处置时间，又可以节约企业人员成本。

【应用效果】

该项系统在部分运营商已投入使用，助力运营商实现了基于 SOAR 的自动化安全运营处置，从监测-分析-研判-编排响应-设备联动-管理执行形成一体化的自动处置流程，安全运营从监测、分析、研判、智能化决策到编排及自动化响应 7*24 小时自适应持续闭环过程；与传统运维方式相比大幅降低 MTTR 分析、响应和处置时间。

步骤	**传统运维**		**自动化运维响应**	
	涉及动作	**时间**	**涉及动作**	**时间**
Step1	告警发现	10 分钟	触发事件告警	20 秒
Step2	事件分析	80 分钟	案例匹配	30 秒
Step3	情报取证	20 分钟	启动自动化编排	30 秒
Step4	发起封堵申请	5 分钟	情报验证	20 秒
Step5	审批员审批	50 分钟	联动处置设备自动对目标 IP 进行封堵	40 秒

Step6	登录设备进行封堵操作	15 分钟	邮件通知相关人员	40 秒
Step7	邮件通知相关人员	20 分钟	N/A	N/A
时间合计	**3 小时 20 分钟**		**3 分钟**	

【下一步工作建议】

1、进一步优化改进机器学习算法，丰富 SOAR 处置体系：通过不断的积累剧本库，来丰富 SOAR 处置体系，为企业安全运营提供了一种全新的模式

2、优化人、技术、流程深度融合机制：变传统的人工监测、预警、分析、响应处置为安全编排及自动化响应处置，提高了企业安全运维效率，降低了企业安全运营成本，助力企业在体系化安全建设大潮中迅速转型。

3.1.6 异常检测

基于 UEBA 技术的用户异常行为监测

【场景描述】

现有安全解决方案对诸如账号使用异常、内部人员盗取信息、特权账号滥用、邮件异常发送、内部用户滥用、数据泄露、主机攻陷、账号失陷、APT 攻击、低频爬虫、撞库攻击等异常检测无能为力。主要困境如下：

- 内部威胁的数据源缺失，无法洞察边界内部安全

包括传统 SIEM 在内的很多安全解决方案，都重度依赖边界流量来提供可见性；即使有威胁情报馈送，也是针对外部，非内部。但很不幸，边界内部也有很多很重要的东西。横向移动、内部应用误用和凭证窃取之类的事件通常都发生在公司内部。然而，公司内部恰恰是很多公司企业都难以获得足够可见性的地方。

- 基于静态规则，容易产生大量误报和噪音

随着接入数据的不断增多，基线也随之动态变化，传统 SIEM 并不具备足够的分析严谨性，会产生大量误报和噪音。安全人员往往从追逐大量误报开始。成熟的公司会学着调整工具，让该软件理解什么是正常事件，以此来降低误报数量。但另一方面，一些安全团队会跳过该步骤，习惯性直接无视太多误报，有可能错过真正威胁。

- 用户行为视角缺失

在完整的组织安全视角，用户视角是非常核心的分析视角，基于用户视角，可以对许多潜在的威胁行为进行有效分析，对于网络中活跃的各类用户及其行为进行精准监控与分析，是对传统资产视角、业务视

角的有力补充。因为黑客一旦侵入系统内，他们便可以伪装成普通用户。而 UEBA 引擎会把注意力放在特定用户的活动上，通过多种统计及机器学习算法建立用户行为模式，当黑客的行为与合法用户出现不同时进行判定并预警。

- 基线、聚类和长周期时序分析等算法能力缺失

只要用过传统 SIEM 调查事件，就会很快发现调查过程磕磕绊绊，对传统 SIEM 系统有限的分析和查询能力感到绝望。今天的调查要求工具拥有足够的灵活性和功能性，需要能结合 AI 算法，对各式各样的大量数据做深入查询。

安恒 AiThink 用户与实体行为分析系统采用的 UEBA 技术（User and Entity Behaviours Analytics），是网络安全领域里发现异常行为是一种重要的能力。很多时候，异常事件多数是一个小概率事件，由于用户对这类事件的精准度要求高，长期以来缺乏有效的检测机制作为保障。在传统检测机制中，我们过分依赖已知威胁检测（譬如 IDS、IPS、NGIDS、NGIPS、FW、NGFW 等）的已知规则来做检测，检测引擎里内置规则或经验，但通过已知规则的机制，规则阈值缺乏灵活性容易引起误判，准确度不够。而 UEBA 手段则是从数据分析的视角去发现关键问题，从聚焦数据内容本身到内容上下文关系、行为分析等，从单点单条检测到多维度大数据分析来发现更多更准确的有价值信息。

Use Case①： 账号盗用

账号盗用一直是困扰企业用户安全审计和行为审计的痛点。恶意用户从事非法活动时，通常会删除或篡改活动日志，从而伪装成其他用户，来掩盖自己的痕迹。

用户 A，是一家金融机构的系统管理员，他的账号有很多出入 IT 系统的权限，某天他的账号被黑客盗用了，黑客通过 VPN 等通道接入内网，并且将数据偷盗到公司外。

UEBA 通过接入的防火墙、IDS 和 IPS 等检测设备日志，发现用户使用从未使用过的外部地址，即源地址行为异常；且利用 LSTM 算法发现用户有大量的数据库查询访问操作，偏离日常行为基线。综合判断用户 A 疑似被账号盗用。

图 1 数据库 SQL 执行返回行数偏离个人基线

Use Case②： APT 攻击

APT 攻击，即高级可持续威胁攻击，指某组织对特定对象展开的持续有效的攻击活动。这种攻击活动具有极强的隐蔽性和针对性，通常会运用受感染的各种介质、供应链和社会工程学等多种手段实施先进的、持久的且有效的威胁和攻击。

某省政府公众服务类网站，攻击者其主要目的是爬取数据并经过二次分析或者加工对外提供有偿性服务信息，攻击者通过伪造 useragent，利用爬虫程序使用超过 500 多个的 c 段 ip，实现多源低频的爬取信息。

从请求数、GET 请求数占比、HTML 请求占比标准差、平均请求发送字节数等角度，使用 UEBA 技术确认攻击源为多源低频团伙爬虫。针对多源低频的攻击行为特征，通过聚类将行为特征放大，并拉长分析的时间轴，往往可以找到攻击团伙深层次的异常行为。采用潜伏型异常检测算法，UEBA 通过长时间轴聚类分析，挖掘深层次异常行为。

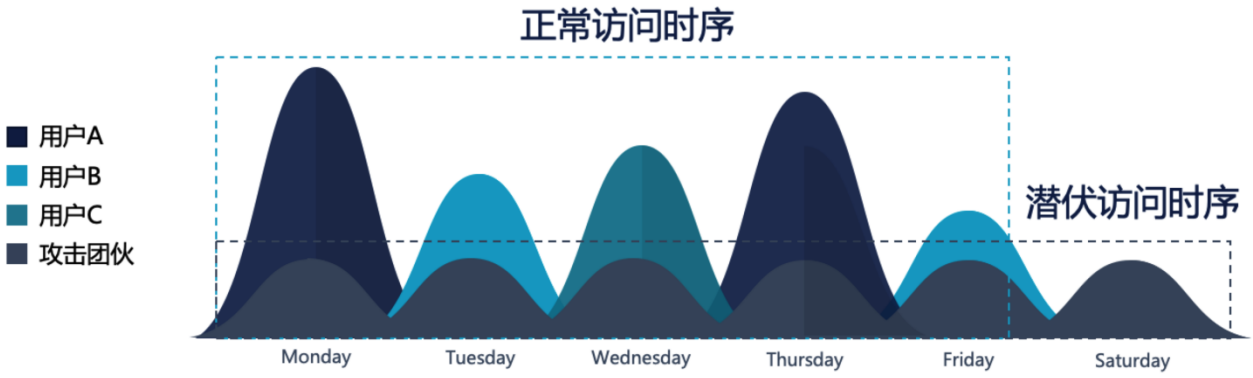

图 2 用户访问时序

【技术方案】

AiThink 采用多种人工智能算法，通过与过去的行为基线或同行群体进行对比，以查看用户或资产行为中的偏差。AiThink 为每个用户、设备、应用程序、特权账户和共享服务账户创建基线，然后检测标准偏差。随后，分配一个分数来指示相关威胁的强度，让企业不仅可以每天查看警报，还可以全时监视顶级恶意用户并采取预防措施。

第一是客观采集人员访问行为，从审计的角度进行统计、展示；

第二是结合用户角色和行为特征，进行内部用户行为画像；通过账号“风险分数”而不是告警的方式，告知运维人员关注异常账户，并作出响应

第三是结合具体场景，通过算法集合、规则、特征等进行多维度的异常行为监控与风险预警。根据客户情况，可以先通过咨询、人工溯源等手段，找到基线后，更针对性的制定安全策略。

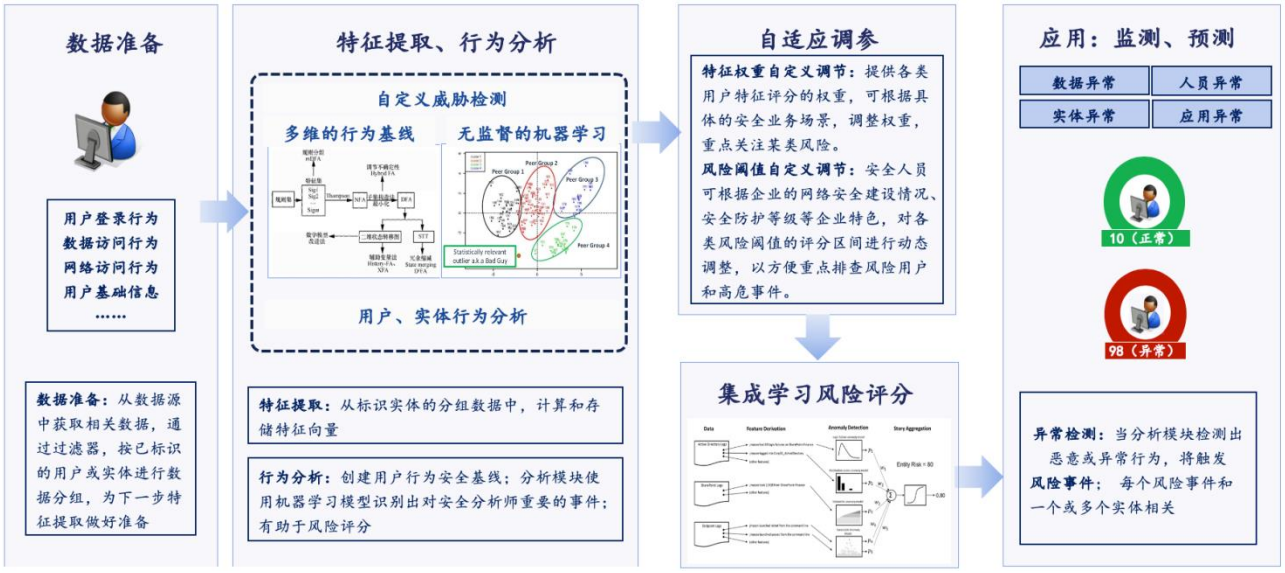

图 3 AiThink 异常检测流程

风险评分

AiThink 采用的 UEBA 技术，把安全运维从事件管理转换到用户、实体风险，极大的降低工作量、提升效率。而这种转换的关键，就在于风险评分。风险评分需要综合各种告警、异常，需要进行群组对比分析，需要考虑历史趋势。风险评分技术还有一个重要的点，在于风险的传导，这需要一套类似 PageRank 的迭代评估机制。风险评分的好坏，会直接影响到 AiThink 实施的成效，直接影响到安全运营的效率。

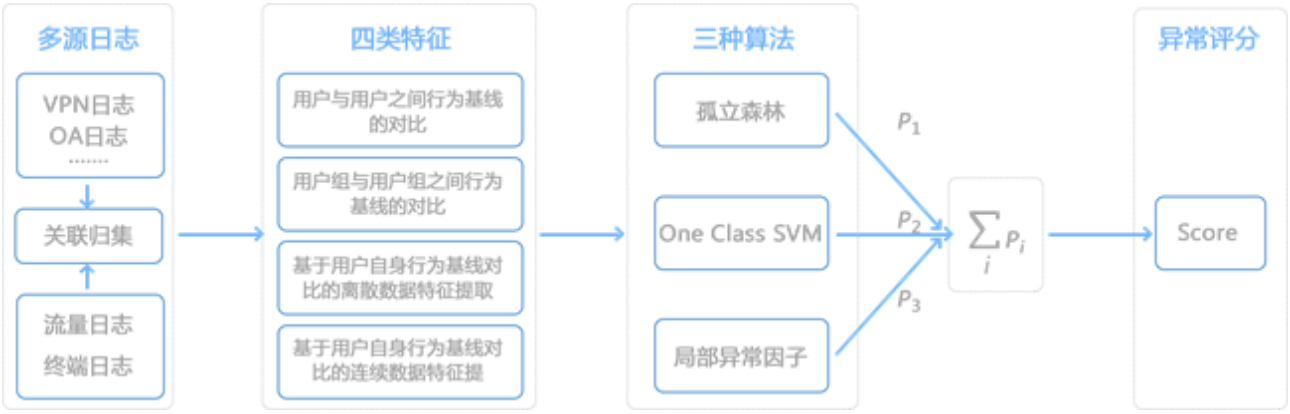

图 4 集成学习风险评分算法

机器学习

AiThink 支持监督式学习、无监督学习。针对一些特定的案例，部分能够支持强化学习或半监督学习。

在网络安全领域，标记数据集的成本非常高。针对标记数据缺乏的现状，某些 UEBA 能够采用了主动学习技术（Active Learning）、自学习（Self Learning），充分发掘标记数据和无标记数据的价值。这部分技术相对来说比较新颖，也有待市场的检验。

Paper: 2018 IEEE（DSC）：A Robust Change-point Detection Method by Eliminating Sparse Noises
《一种基于淘汰稀疏噪点的时间序列异常点检测方法》
Paper: 2018 IEEE（DSC）：A Categorically Reweighted Feature Extraction Method for Anomaly Detection
《异常检测的范畴再加权特征提取方法》

论文：基于机器学习的用户实体行为分析技术在账号异常检测中的应用[J].通信技术，2020，53（05）：1262-1267.

Patent: A network traffic anomaly detection method and system
Patent number: 201710803213.1
专利：基于集成学习的异常用户检测方法及系统 专利号：201910751220.0

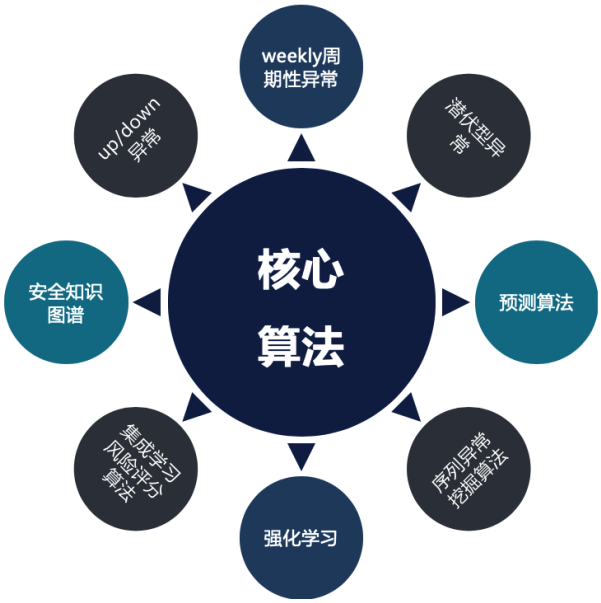

图 5 核心算法

图 6 机器学习算法

知识图谱

知识图谱已经成为人工智能领域的热门领域，在网络安全中也有巨大的应用潜力。安恒 AiThink 支持安全知识图谱能力，可以把从事件、告警、异常、访问中抽取出实体及实体间关系，构建一张网络图谱。

任何一个事件、告警、异常，都可以放到这个网络图谱中，直观、明晰的看到多层关系，可以让分析抵达更远的边界，触达更隐蔽的联系，揭露出哪怕最细微的线索。

结合攻击链，知识图谱的关系回放，还能够让安全分析师，近似真实的复现攻击全过程，了解攻击的路径与脆弱点，评估潜在的受影响资产，从而更好的进行应急响应与改进。

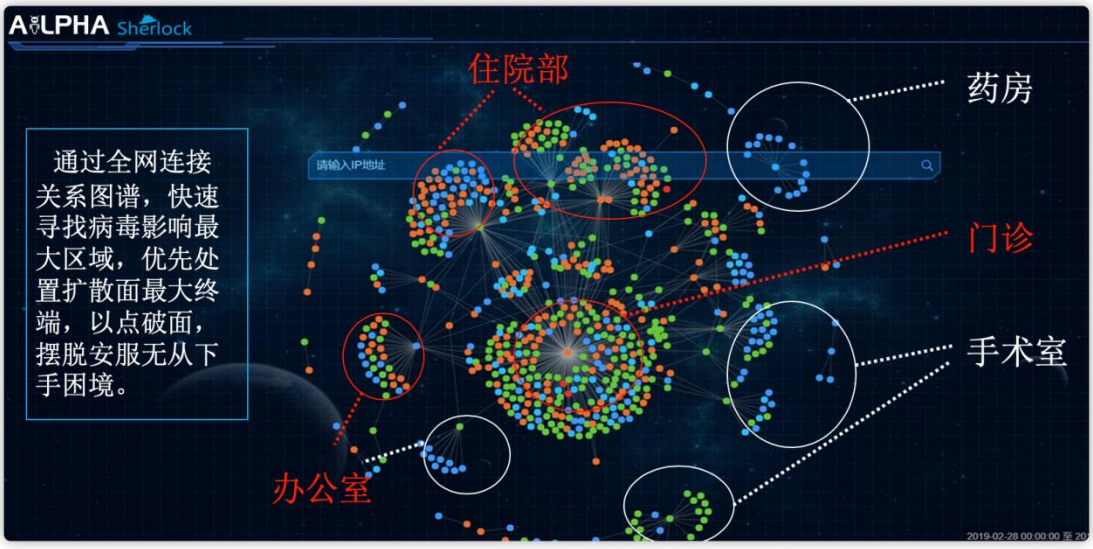

图 7 知识图谱

【应用效果】

AiThink 可以用于行为分析覆盖的多个应用领域，可以帮助用户应对内部威胁、账号失陷、恶意内部人员、数据泄露、身份与访问管理、特权账号滥用和误用等众多领域。

某上市企业虽然已部署了流量分析和审计类的一些安全产品，但对于账号风险检测一直存在缺失，经常出现员工已离职却还在下载文件等安全问题。内部一共部署 6 台大数据节点，总数据量 3PB，日增原始数据 8TB 左右。每天支持 100 亿的事件和历史总记录超过 3 万亿条记录，这些记录都存有原始数据（非聚合数据），随时可以使用 SQL 查询和分析，生成用户报告。迄今为止已发现上百个安全风险，包括非法外连、VPN 账号共享和离职员工账号未清除等。

【下一步工作建议】

AiThink 采用的 UEBA 技术，将转向具有高级分析功能的“现代安全信息和事件管理”（Modern SIEM）系统，或嵌入 UEBA 功能的其他工具。SIEM 在分析方面变得更好，可以提供更复杂的用例，同时，AiThink/UEBA 专注于更好的数据管理和可操作性，这使得他们更接近于 SIEM。目前，各大安全厂商也在积极采用并购或自研发的方式，将 UEBA 的功能嵌入到各自的系统中，UEBA 技术的运用必将极大提升安全事件的检测能力。

数据库安全访问

【场景描述】

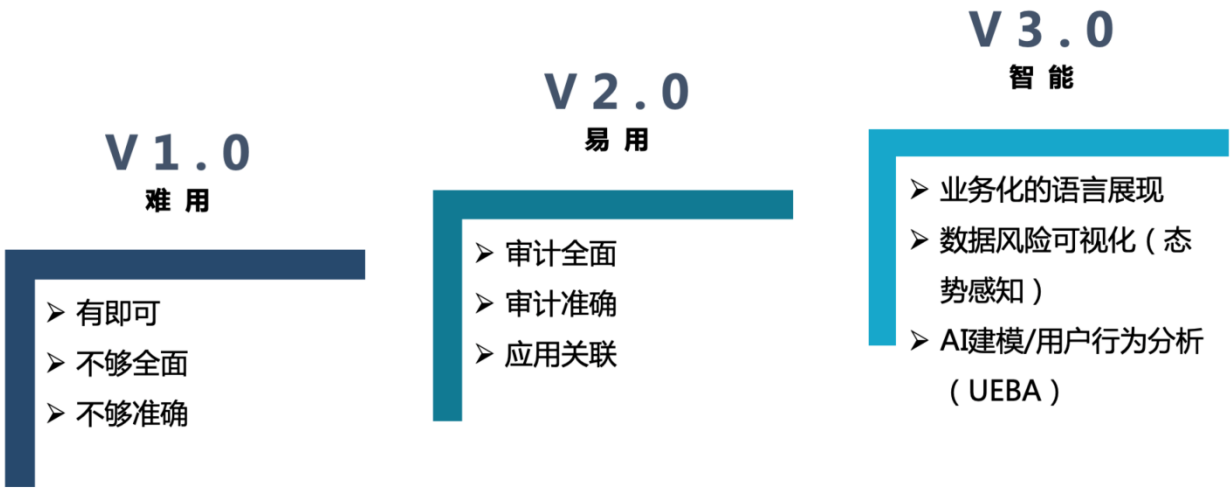

图 1 数据审计系统前世今生

一代

系统能记录对数据库的访问，且审计结果可查询并展现；而对于审计全面性、准确性的要求则比较简单，有时甚至不太关注是否能够做到全面审计。更有甚者，有的厂商基于传统的网络审计产品简单改造或产品都未经改造只是概念包装后就推向市场的产品。招标参数，结果列出一堆网络审计产品要求：支持包括但不限于 HTTP、POP/POP3、SMTP、TELNET、FTP 等。

二代

审计全面

1. 审计的内容全：能够全面记录会话和语句信息等；

2. 兼容的数据库类型全：能够兼容各种主流的关系型、非关系型数据库（NoSQL），以及大数据平台组件等。

审计准确

审计产品通过 DPI 技术对各种数据库通讯协议进行分析，以还原通讯包中的通讯协议结构，继而准确识别 SQL 语句、SQL 句柄、参数、字符集等信息；通过‘会话’、‘语句’、‘风险’之间的内在联系实现界面交互和线索关联，从而提升“审计追踪”的能力和便利性。

应用关联

从审计的角度，由于一个业务系统往往公用一个数据库用户，因此无法区分哪个业务人员触发了哪些数据库操作，因此不能真正地满足追查的需要。为了满足这个需要，部分厂商通过时间戳方式关联，然而关联审计信息并不准确，尤其在高并发场景中，正确率不超过 50%；部分专业的厂商如安恒，基于插件通过 http 协议与数据库协议进行关联，实现 100%的应用关联审计。

三代

业务化的语言展现

三代数据库安全审计产品统计和追踪按照业务的行为和分类来进行信息的组织和展现。不仅仅是审计记录的展现，还包括数据的组织，把分散的 SQL 语句，再组织成一个个业务操作，这个时候给业务人员展现的就不是每秒有多少个 SQL 操作，而是每秒有多少个业务操作。当前一个会话中的多条审计记录，组织成一个业务操作后，不仅仅是审计记录的展现，包括性能、类型统计、成功与失败、检索条件、报表等都是基于业务操作为单位。

数据风险可视化（态势感知）

不仅要知道“数据被谁访问”，还需要知道“谁访问了哪些数据”，用户行为分析视角成为了一个全新的数据库审计视角。借此更加智能化的满足合规运营、风险事件监测和风险趋势分析等需求。

AI 建模/用户行为分析（ UEBA ）

数据库审计系统的“风险监控能力”，是企业安全部门关注的重点，包括是否存在对数据资产的攻击、口令猜测、数据泄露、第三方违规操作、不明访问来源等安全风险。因此，系统需要支持更加全面、灵活的策略规则配置，准确的规则触发与及时告警的能力，从而在第一时间发现并解决风险问题，避免事件规模及危害的进一步扩大。此外，数据库审计系统应具备自动化、智能化的学习能力，通过对重要数据资产

访问来源和访问行为等的“学习”，建立起访问来源和访问行为的基线，并以此作为进一步发现异常访问和异常行为的基础。

Use Case①：数据泄露

AiThink 基于用户访问数据自动生成安全基线。比如针对访问数据库的行为，从谁访问的、如何访问、访问了什么数据、同一类型的用户访问行为是不是一致、访问量多少、什么时间访问等多个维度来构建动态的基线。这比传统的基于规则或者基于阈值的检测大大提升了检测的准确度。

举例来说，一些客户核心业务系统中防止数据泄漏的做法，是针对一些营业人员的高频次访问客户详单这种敏感操作做安全审计，采取了一刀切的审计策略。比如：营业人员如果一天内上千次的查客户详单，认为会涉及到数据泄露风险。但是每个公司下面各个营业点或者各个子公司的业务量是不一样的，

业务有繁忙和空闲周期，这里面就没办法有效的查出真正的“有数据泄漏”的点，反而有很多误报。AiThink 可以基于动态基线，结合每个人的操作历史行为、登录地等特征，再结合历史操作的量来对今日的操作行为进行判断，这样的一个检测效果就会比原来检测的准确率提升很多。

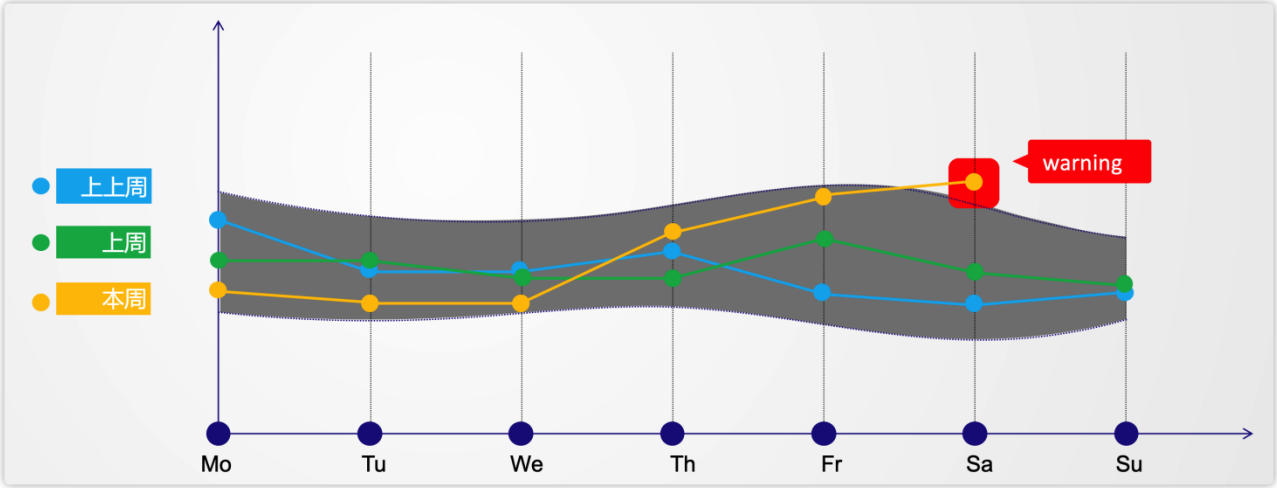

图 2 历史基线置信区间

Use Case②： SQL 注入

在一些特点的生产系统，进入业务稳定期后，功能模块相对稳定，具体到每个访问行为，其产生的 SQL 语句模板也相对固定。反之，一个 SQL 语句模板被业务系统请求的来源也相对固定，可见，在一个稳定运行的业务系统中，应用访问行为与 SQL 语句模板存在相对稳定的对应关系；无论应用访问行为出现了新的语句模板，或者产生的语句模板出现了新的应用请求来源，都可能成为可疑的访问行为。

比较典型的应用场景：SQL 注入伪装成正常的数据访问，但 AiThink 在行为模型中发现业务系统应用请求中出现了新的语句，产生疑似非法操作的告警，从根本上解决 SQL 注入的风险。

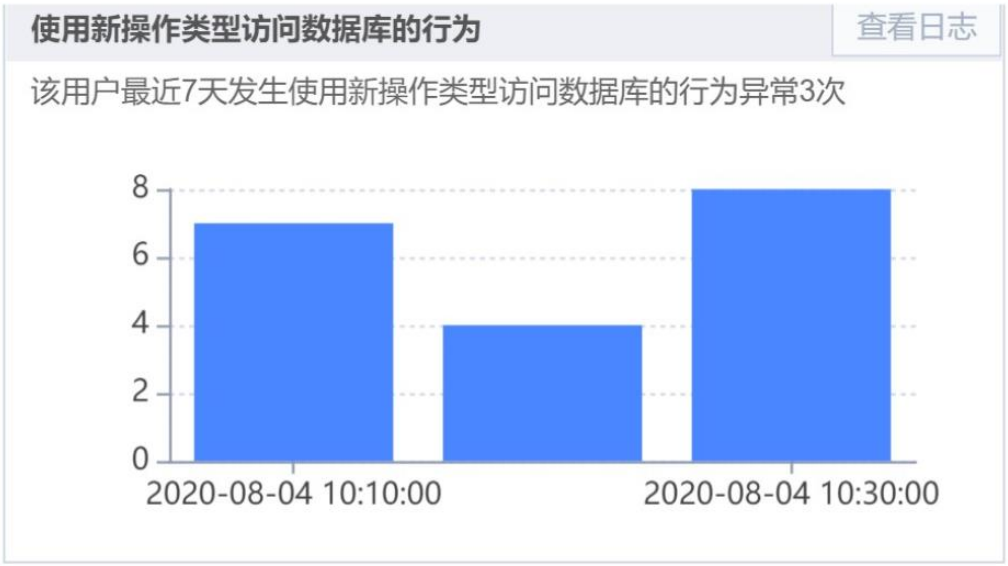

图 3 使用新操作类型访问数据库的行为

Use Case③：特权账号监控

主机登录算法通过分析每个用户经常使用的主机或者服务来识别异常。例如用户 A 通常倾向于登录固定的几台机器，并且都是使用相似的命令，而销售人员都是使用 OA 系统的服务进行数据访问，两者使用方式完全不同。若发现用户 A 登录了一台保存销售数据的服务器，这与他所在的 DBA 管理员组群体行为不同，则可能用户 A 存在异常。

AiThink 通过 Kmeans 聚类算法根据用户行为数据的特征矩阵对用户划分对等组，行为模式类似的人群会划分到一个动态群组。

基于 Peer Group Analysis 进行异常行为识别。异常用户一般占少数，可以通过对大部分用户的行为进行建模，找出少数的高危用户，再匹配威胁或攻击模型来确认。

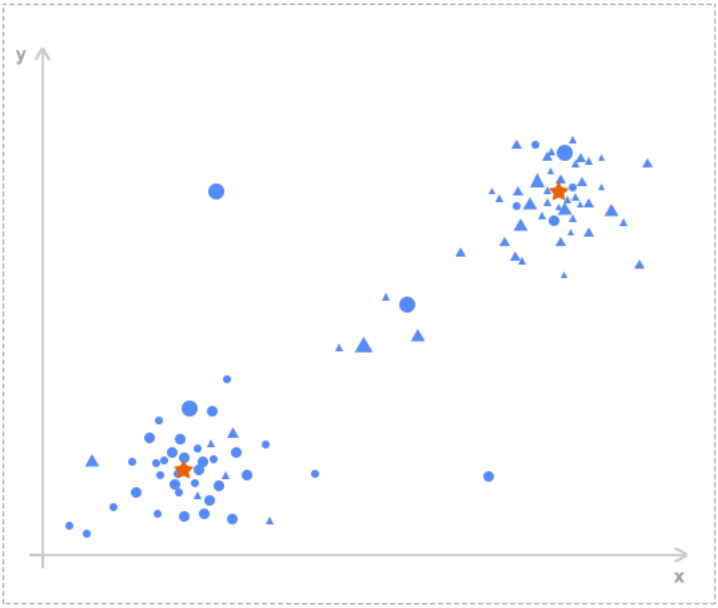

图 4 聚类算法

Use Case④：敏感表监控

从操作类型角度，表对象经常会以什么样的方式被访问，访问频次怎样；从语句模板角度，表对象经常会被什么样的语句模板来访问。

例如某表对象出现了新的访问类型（如经常 SELECT 的表对象突然出现了 UPDATE 的操作行为），或者表对象出现了一种新的语句访问模型（如出现了一种新的语句模板），疑似为非正常访问动作。

某电商公司对客户实行积分制，客户在公司网站上购物后可以积分，在积累一定的积分之后可利用积分兑换人民币，并用绑定的银行卡提现。程序是待客户通过积分向公司申请提现后，公司后台自动审核客户的资料，一旦通过便自动将资金转到客户指定的银行账户，并设置了转款限额为 3 万元。

A 某和 B 某团伙作案，A 某负责操作电脑攻破公司网站安全防护并修改账户，然后将其网站内的资金盗转出来，B 某负责提供黑卡并洗钱。

A 某进入公司后台服务器后，利用 B 某提供的黑卡进入网站数据库更改了客户本身绑定的银行卡，又更改了该客户对应的积分，然后申请提现，由于控制了后台服务器，可以随意更改设置，操作很顺利，两人一共盗转了 320 多万。

AiThink 通过敏感表监控，发现该客户积分数据表中出现了新的 update 操作（正常情况下新增购物记录积累积分，该客户积分数据表都是 insert 的方式插入数据行来增加积分），A 某为了方便操作，直接对以前旧购物记录行进行了 update 积分值操作，即表对象出现了新的访问类型，AiThink 对敏感表的监控发现异常，最终确认并抓获该犯罪团伙。

【技术方案】

AiThink 数据库访问安全解决方案，能够对进出核心数据库的访问流量进行数据报文字段级的解析操作，完全还原出操作的细节，并给出详尽的操作返回结果，通过内建的机器学习 AI 引擎，使用机器学习算法来确定用户和数据库行为基准，以检测异常。

从客户的时间维度来看，数据库的访问是有规律的，客户的业务时间也是有规律的。AiThink 可以根据用户历史访问活动的信息刻画出一个数据的访问“基线”，而之后则可以利用这个基线对后续的访问活动做进一步的判别。

举例来说，他能够识别在非工作时间访问敏感数据的授权特权账户，或运行过多偏离标准查询的特权用户账户。他将这类特征场景与现有策略及其他指标相结合，利用集成学习风险评分算法，计算出用户综合风险评分。如此一来，安全分析师便可调查特定的风险用户详情并采取措施解决潜在的内部威胁。借助基于数据库和数据库账号的钳形可视化视角，安全团队可以了解关键数据的位置、他们的访问者以及面临的风险。您可以通过简单、直观的用户体验保护您的数据库访问安全。

图 5 技术方案亮点

自动识别可疑活动

AiThink 数据库访问安全解决方案具备自动化、智能化的学习能力，通过对重要数据资产访问来源和访问行为等的“学习”，建立起访问来源和访问行为的基线，并以此作为进一步发现异常访问和异常行为的基础。

比如以敏感对象表元素频度访问、敏感对象的操作类型、访问源的行为、去参数化的语句模板、去参数化的应用 URL 行为等，通过一定组合方式，形成多样的数据分析模型，从而自动发现危险的用户数据访问行为。

钳形战速视角

➢ 监督所有数据库活动

获得企业范围内所有数据库事务的可见性，包括本地特权用户访问和服务账户活动。AiThink 数据库访问安全解决方案持续监控系统环境，并收集所有登录/注销、更新、特权活动等的合并记录，以创建细粒度的审计跟踪，为每个数据库指明“谁，何时，何地以及如何，做了什么”。

➢ 监督所有用户活动

传统数据库审计产品，以“数据被谁访问”的视角来反向溯源，审计出来的数据都是以数据库为单位，“谁访问了哪些数据”成为一个全新的数据库审计视角，这种以应用发起者为出发点，通过哪些应用业务的请求，最终访问了哪些数据，用户行为审计视角是非常关键核心的分析视角。

图 6 数据安全态势感知可视化

包含 DB 账号和数据库等近 30 种的数据源和风险类型视角，便于用户全方位洞察系统安全情况，精准定位数据泄露风险。

针对已定位出的特定风险用户或者数据库，进行特定行为画像，可以全方位展示该用户或数据库的全局画像信息，特征对应事件快速自适应排序，通过风险趋势图以及多种可视化图表，便于用户排查最有风险事件。

通过对等组分析识别可疑数据访问

通过数据库活动信息分析特定用户的数据访问行为。调查特定于个人的事件和异常，查看典型用户活动的基准，并将给定用户与该用户的对等组进行比较。

通过构建群组分析，可以跨越单个用户、实体的局限，看到更大的事实。通过对比群组，可以更容易检测出异常，更重要的是，通过综合研判，可以降低误报，提升信噪比。

关注监控特权帐户

使用共享帐户时，往往很难知道确切是谁在访问系统和数据。如果有人添加后门帐户，进行未经授权的更改或执行其他一些存在风险的操作，很难跟踪到哪些人员执行了哪些操作。

借助 AiThink，可以针对特权账号添加关注，为审计和取证做好全面准备。

加速违规调查和响应

界面简单易用、操作方便。提供 SQL 语句翻译能力，帮助非技术人员快速了解数据库访问情况。用简单的语言解释紧急事件。您不必是数据库专家就可以进行成功的调查。AiThink 提供的用户与数据库风险分析使您可以细粒度地了解谁使用数据的方式，并提供可行的见解（描述、详情和处置建议），因此您可以在损坏发生之前迅速遏制数据泄露。

开箱即用

基于 13 年国家级网络安保经验，内置 30+UEBA 特征场景，用户开箱即用。立即保护，快速实现价值。

通过预置的安全策略快速部署，以阻止数据库攻击威胁，并可以自动发现以查找敏感数据和隐藏的漏洞。它在协议和 OS 级别上查找威胁和攻击，以及未经授权的 SQL 活动，然后发出警报，并在适当时阻止未经授权的活动以保护数据。

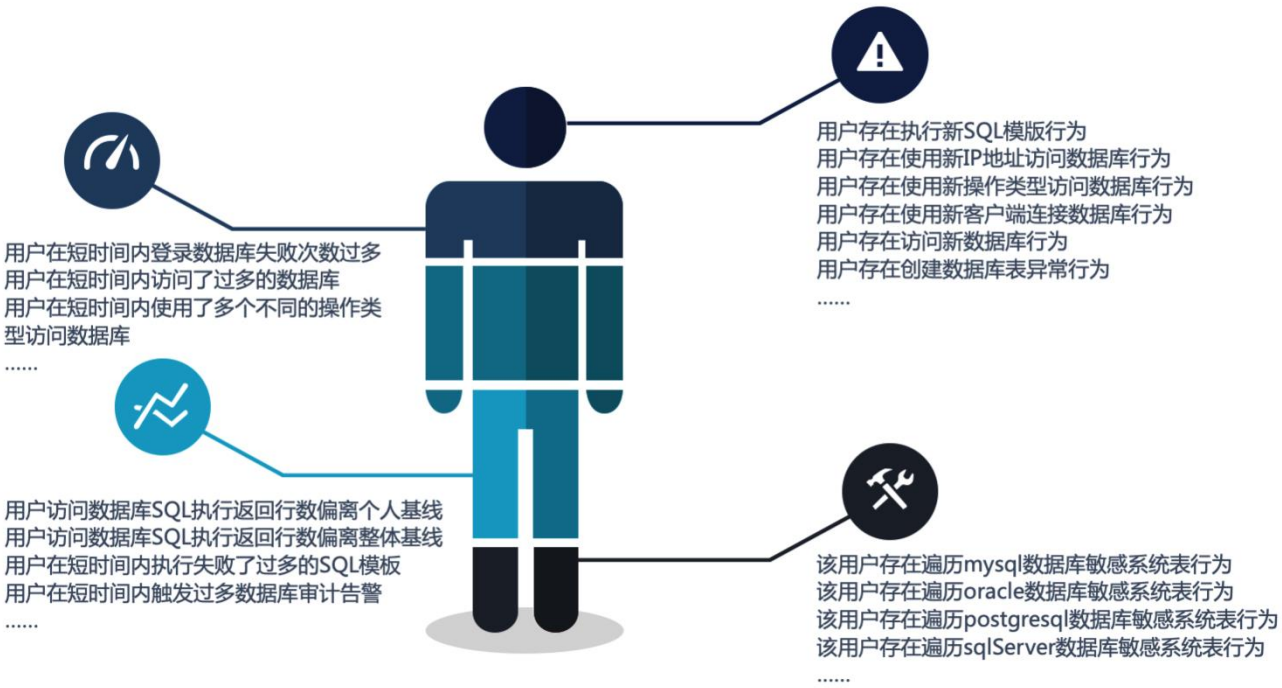

图 7 数据库访问安全特征场景

【应用效果】

通过长时间、持续性地对用户的行为进行记录和分析，根据历史行为分析来检测当前的一些操作是否存在异常，这样就能大大削减告警的数量，能够迅速地关注到存在的风险点。该方案已帮助众多企业组织发现多起数据泄露风险，诸如“遍历 Oracle 数据库系统敏感表行为”、“静默数据库账号批量修改数据库密码字段值”和“新出现用户通过已有数据库账号异常操作行为”等众多异常行为，大大降低了其数据泄露安全风险。

【下一步工作建议】

正如互联网预言大神凯文·凯利在对未来趋势的总结中所言：“人类正处在一个数据流动的时代，数据的重要性空前提升且不断发展，处理数据已变得同处理客户一样重要。”事实上，只要想想人们现在每天要花费多少时间用手机浏览各种图文和视频信息，每分钟有多少笔在线支付或转账在发生，还有那炙手可热、野蛮生长的区块链和比特币等等…不难发现，人类社会的知识、财富乃至历史正在经由数据承载，而由此产生的、源源不断的数据，又进一步推动着商业模式创新的车轮滚滚向前，数字化已渗透至社会的各个角落。

数据的安全不仅需要如上述提及的基于 UEBA 能力结合分析的数据库访问安全保障，还需要从数据分级分类、脱敏、加密、水印和 DLP 等多角度全方位进行数据安全保障。

5G 网络终端异常检测

【场景描述】

网络运维面对海量监控曲线，如何在其中快速发现异常，从而及时响应？在运维监控领域，伴随着云化、大数据、IoT 等业务的发展，被管对象成指数级增加，对于海量的监控指标，需要有更加智能化的异常检测方案，从而实现快速、准确、低成本的监控预警。

华为依托长期服务运营商的经验和研发优势，开发了人工智能流量异常检测方案，通过人工智能的帮助，分析历史数据，拟合指标趋势，从而实现指标异常检测，进而支持触发告警、人工标注反馈，完成异常检测和处理的闭环。

下面，以华为针对海量终端可能引起的信令风暴的威胁检测方案为例来说明信令风暴威胁检测功能。异常检测区分不同场景，例如，针对网元状态异常检测，漫游信令异常检测，及信令风暴异常检测等。，华为 5G 网络终端异常检测方案可支持以下场景的异常检测功能：

- 大量非法身份终端反复接入网络
- 大量合法终端由于业务异常事件反复接入网络
- 大量合法终端被黑客控制反复开机，反复退网/入网
- 大量合法终端被黑客控制周期性发送信令，单个 UE 慢速发送信令攻击

- 大量合法终端被黑客控制，发送大量流量浪费网络资源
- 大量合法终端被黑客控制，扫描和入侵其他终端和网元

【技术方案】

业务流程和功能要求：

5G 安全检测引擎工作业务流程，利用大数据，基于人工智能分析和检测信令协议状态机变化特征，精准判定异常 UE 和恶意 UE，并且可以提供闭环处置的手段，整个流程如下图所示：

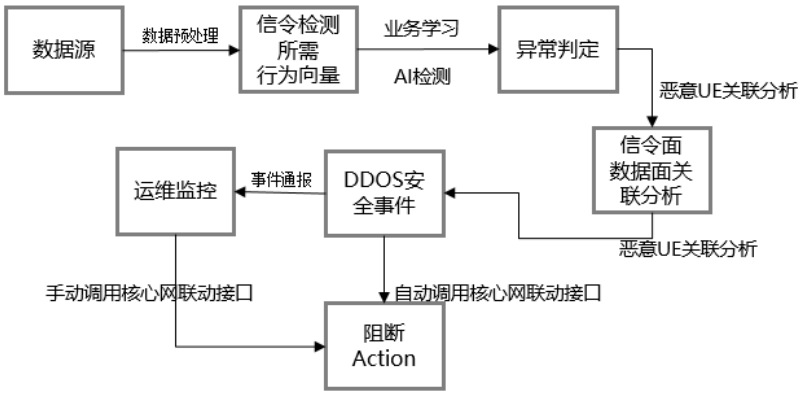

图 1 5G 信令风暴检测场景业务流程图

一、数据采集和行为建模

1. UE 在无线网络和核心网络内部的信令日志（以协议状态机变化为基准）

a) 无线侧 UE 协议状态机行为数据

- UE 协议状态机变化事件总数
- UE 协议状态机变化单项事件计数（RRC 建联和重建等）
- UE 协议状态机变化序列

b) 核心网侧 UE 协议状态机行为数据

- UE 接入时间段
- UE 在制定时间段的状态机变化次数
- UE 协议状态机事件单项事件计数（附着和去附着等）
- UE 协议状态机变化序列
- UE 数据面数据总量，速度，包长等特征

2. 数据获取形式

a) 在线实时流量和日志

b)　离线带有时间戳的流量和日志

基于在线和离线的数据和日志提取关键特征，通过人工智能训练得出正常业务的行为向量。有了正常的 UE 行为向量模型，就可以发现异常 UE 的群体集，如下图所示

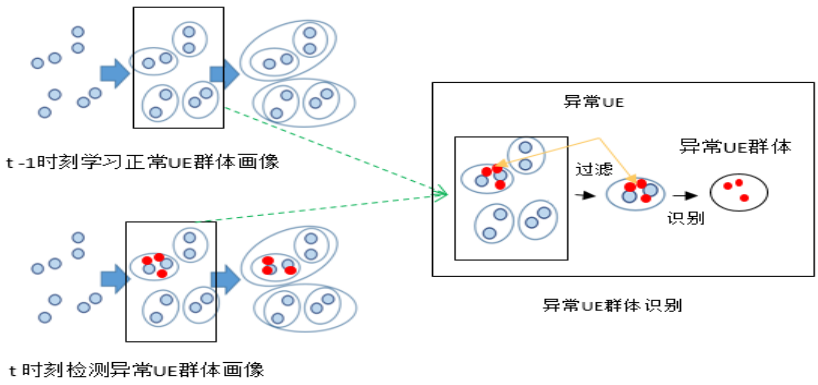

图 2　UE 行为基线学习和对比分析方式示意图

二、异常检测

5G 安全检测引擎可收集大量实时数据或者离线数据，提取关键特征作为正常业务的行为模型，并作为正常向量；或者通过已知攻击的流量样本中提取攻击模型，并作为异常向量，将正常向量和异常向量导入 5G 安全检测引擎。

5G 安全检测引擎可以从实时数据源中提取当前时刻的行为向量，通过与已知的正常向量和异常向量对比，判定是否有异常群体 UE。

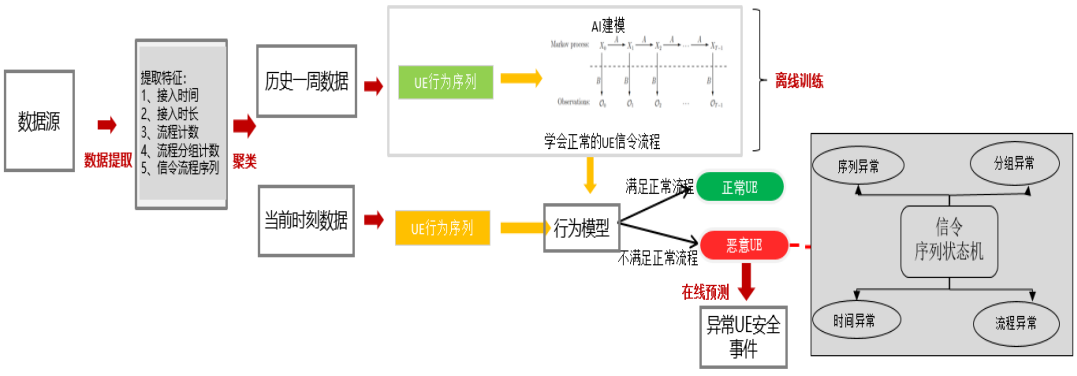

图 3 网元异常检测流程示意图

三、关联分析

仅仅凭借单维度信令面人工智能分析，在模型积累初期可能会遇到检测率不准的情况，5G 安全检测引擎应针对信令域异常的 UE 进行数据分析，关联信令域和数据分析结果，判定是否为恶意 UE，并且提取 UE 的 IMSI，可供威胁闭环处理。

数据域分析包含：

- 外联数据检测
- 异常扫描检测

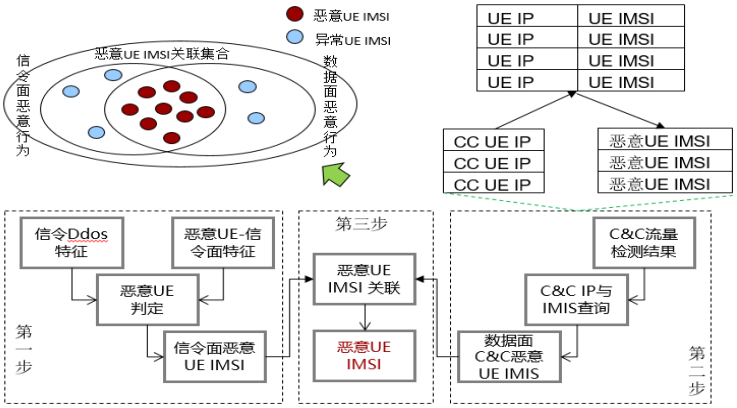

图 4 关联分析流程示意图

【应用效果】

异常阻断

5G 安全检测引擎应可以针对不同类型的检测结果，进行差异化阻断和限速。包括，

- 对于异常/恶意 UE 做限速
- 对于异常/恶意 UE 去附着
- 对于异常/恶意 UE 加黑名单
- 对于异常/恶意 UE 外联通道在 Gi 防火墙侧做限制

5G 智能安全检测引擎，可以通过人工智能和机器学习等技术的帮助，发现人力无法感知的异常行为和攻击特征，预判预防网络攻击的发生，减少因网络攻击带来的损失。目前华为的该智能安全检测引擎已在运营商网络部署，运行正常，获得实际效果。

【下一步工作建议】

信令风暴的异常检测只是智能安全检测引擎功能的一部分，未来将继续扩展人工智能技术的应用，提高人工智能模型分析判断的准确率和效率，提高在有对抗样本攻击环境下的鲁棒性。

随着 5G 网络的大规模部署，5G 智能安全检测引擎也会配合部署在网络的重要位置，为 5G 网络安全保驾护航。

3.1.7 威胁情报

基于实体命名识别技术构建漏洞知识图谱

【场景描述】

在信息安全领域，由各国官方收集维护的缺陷/漏洞数据库是十分权威并准确的威胁情报来源。如美国政府维护的 NVD（National Vulnerability Database）和中国的 CNNVD（China National Vulnerability Database of Information Security）。但单个的漏洞报告信息是零碎的，片面的。对此我们可以利用人工智能的手段，基于漏洞报告构建知识图谱，将孤立的情报关联起来，为用户呈现关于缺陷/漏洞的全景图，如下图所示：

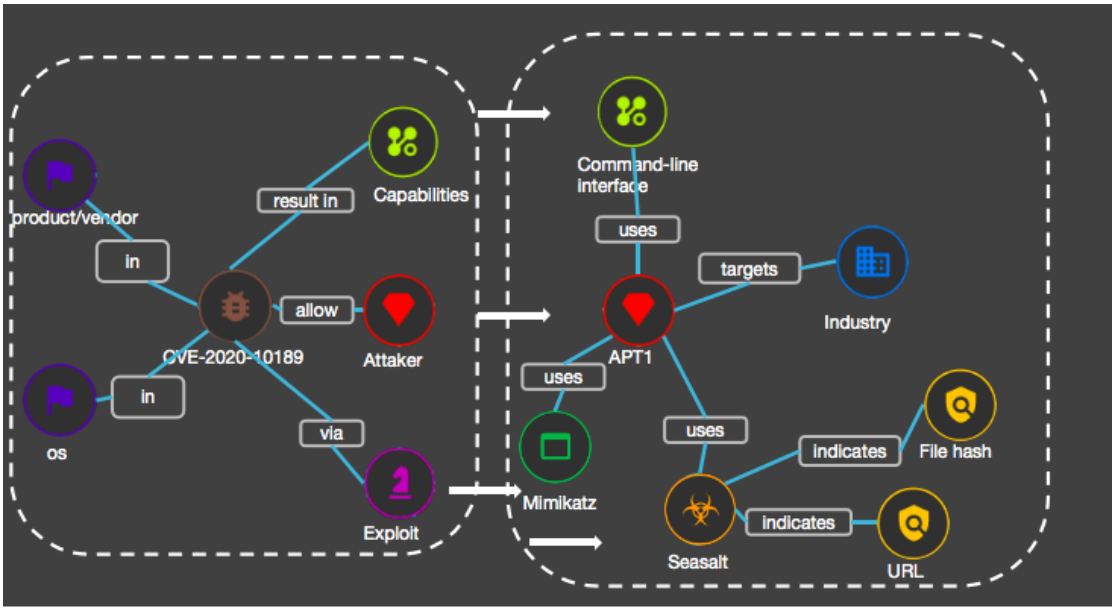

图 1 漏洞知识图谱

在漏洞知识图谱中，我们以漏洞报告中提及的各类实体作为图中的点，实体间的各类关系作为图中的边。用户可以从某一个实体出发，例如某一个 APT 组织，便可得知该组织惯用的攻击模式和攻击工具。这些信息往往是从单个报告中无法获得的，而知识谱图则可以将众多漏洞报告有机的串联起来。

在此基础上，我们还可以利用缺陷/漏洞知识图谱对其他类型的威胁情报进行提取整合，包括 CAPEC 和 ATT&CK、勒索事件、挖矿事件、数据外泄和钓鱼事件等，以及进行新知识的推演，包括属性推演和关系推演。

【技术方案】

一．技术方案概述

亚信安全基于人工智能构建威胁知识图谱的设计思路，是利用利用 NLP 中的命名实体识别（Named Entity Recognition，NER）任务完成实体的提取。整体系统架构如下图所示：

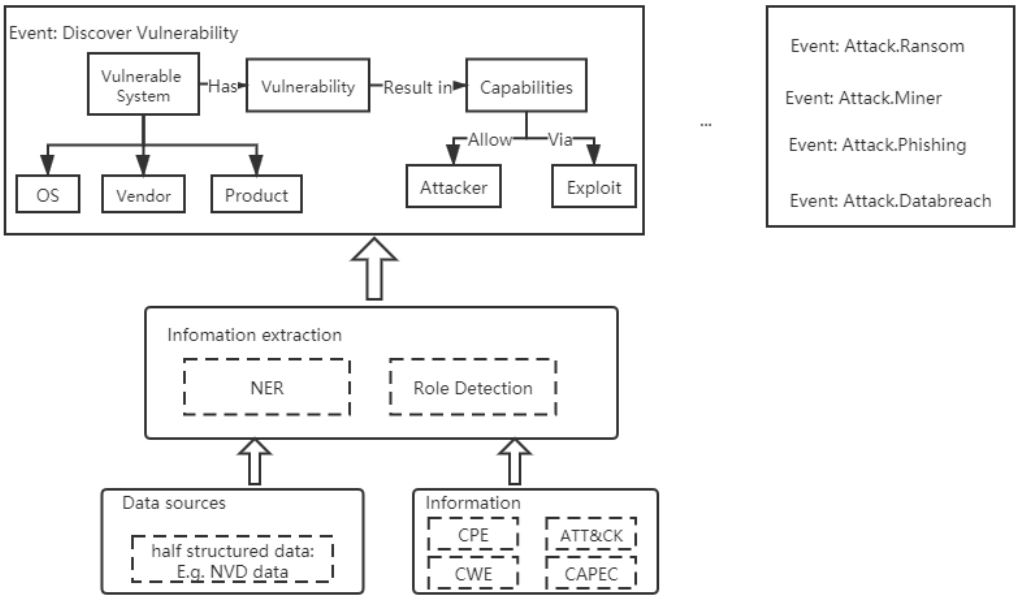

图 2 NVD-NER 架构图

通过输入半结构化的 NVD 数据及其他相关威胁情报，利用实体识别算法和角色检测提取其中的实体并赋予其一定的角色。

实体与角色的对应关系会根据事件场景的不同而有一些差别，但是核心角色应包括：

威胁主体：攻击发起者

威胁客体：受到攻击的目标客体/资产，具体可细分为系统、提供商、产品等。

攻击：包括攻击模式，攻击者所使用的策略技术，在漏洞场景中表现为一个漏洞利用的过程。

隐患/弱点：在漏洞场景中表现为系统的缺陷，漏洞的类型等，可参考 CWE。

攻击工具：攻击所关联的文件、恶意代码，特定方式等，在漏洞场景中对应漏洞利用的方式。

其中，NER 的技术模型如下图所示：

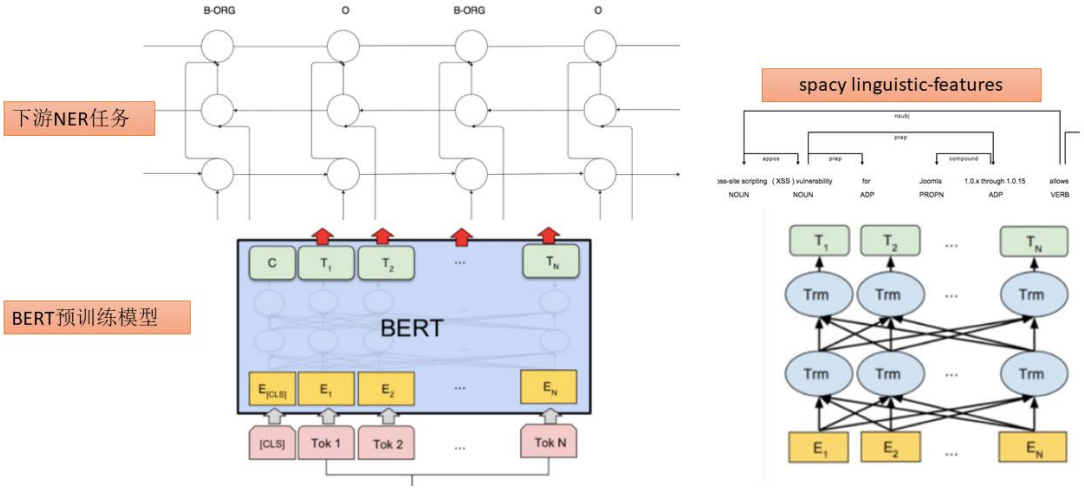

图 3 亚信安全 NVD-NER 模型

首先基于 BERT 预训练模型处理文本，再交给下游 NER 任务，同时依赖于 spacy 工具的基础自然语言特征。

二实体关系构建

关系的构建包括直接关系与间接关系构建。直接关系比较容易得到，内网环境中通常能通过日志、沙箱、原始流量和外部数据直接得到的关系对，例如，文件访问域名，域名解析 IP，文件访问 IP 等。

间接关系是通过间接关联得到的关系，比如使用同一种攻击工具的攻击者有一定的相似性，文件与文件通过相似度计算得到的相似性等等都属于间接关系。这样通过直接关系与间接关系的构建就构成了内网安全知识图谱。

针对信息安全领域知识图谱构建的两个关键要素，构建了威胁元语言模型对威胁知识的结构化描述，包括概念、实体、属性的定义以及知识关系的定义。研究中依据 STIX2.0 以及领域专家知识，构建三层安全知识图谱，如下图所示：

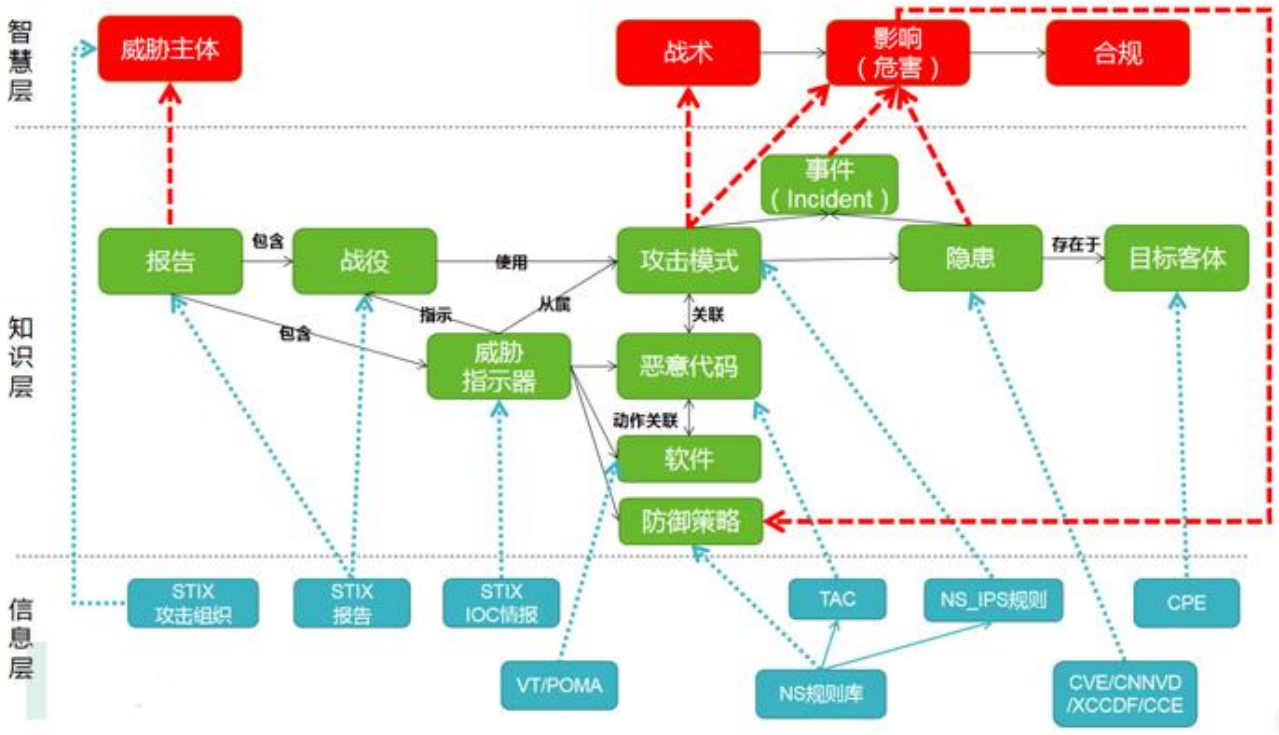

图 4 安全知识图谱

【应用效果】

1. 根据实验结果，本方案的实体提取准确率在 90%以上。分析一篇漏洞报告的时间在分钟级。
2. 该方案已应用于亚信安全超洞察威胁情报平台。
3. 通过威胁图谱提升了对 APT 组织的分析追踪能力，新挖掘出的攻击手段和攻击工具占总量的 50%。

【下一步工作建议】

1. 针对其它类型的威胁情报定义更多的实体类型。
2. 支持分析其它类型的威胁情报，并从中提取实体。
3. 支持自动识别实体间的关系。
4. 持续利用威胁图谱进行 APT 组织追踪，利用各类图算法挖掘更多潜在的知识和规律。

威胁样本筛选、标注与相似样本自动识别

【场景描述】

近年来，网络攻击愈发频繁，其中鱼叉式钓鱼攻击占据非常大的比例。这种攻击的成本低，攻击者可以自建或者使用肉鸡构建邮件服务器，只需要得知单位/机构的负责人或者相关人员的邮箱就可以发动攻

击，同时只要主题选取的好，攻击成功的几率较高。现阶段市场上的对于邮件附件的检测方式基本是使用动态沙箱运行，结合沙箱规则进行告警，并根据一些家族告警来对样本进行标定。但是这种方式具有一些不足之处：一方面，沙箱性能掣肘，沙箱获取样本行为需要一定时间，面对海量样本，则需要海量的设备；另一方面，沙箱对抗成本高。

对于以上问题，启明星辰发挥创新、研发优势，设计研发了一套样本筛选、标注和自动识别相似样本的系统。该系统通过使用不同来源的数据和不同维度的算法，能够筛选出 APT 组织投放的高价值样本。

【技术方案概述】

较为完整的系统结构如下。

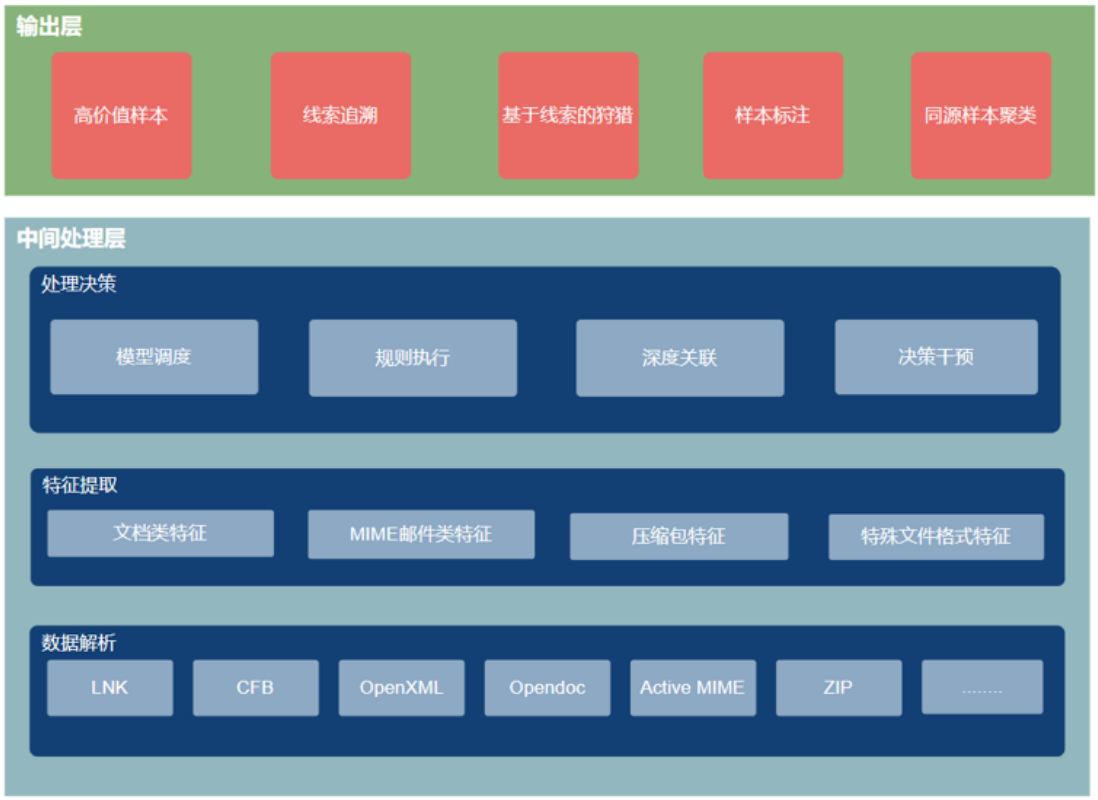

图 1 样本筛选、标注与相似样本自动识别系统结构图

该系统包含如下部分：

（1）数据解析，主要用于提取样本的元数据、代码信息、内嵌文件信息等多维度信息，并进行较小颗粒度的预检测以完成初步判断与筛选。

（2）特征提取，基于数据解析后得到的样本相关数据对其进行特征抽取。

（3）模型调度，采用多模型干预的方式，有层级的进行筛选判断，对于不同类型的样本会选用不同的机器学习模型对其进行黑白鉴定，之后再对黑样本进行特殊性识别等。

（4）规则执行，用于自定义的规则识别与对样本的快速标定。

（5）深度关联，基于静态特征、提取的线索等数据计算样本间同源哈希，用于快速深度关联同源样本并进行标注。

（6）某些数据会基于分析人员的研究进行最终结果的决策干预。

下面以文档类样本为例，描述其处理的过程。

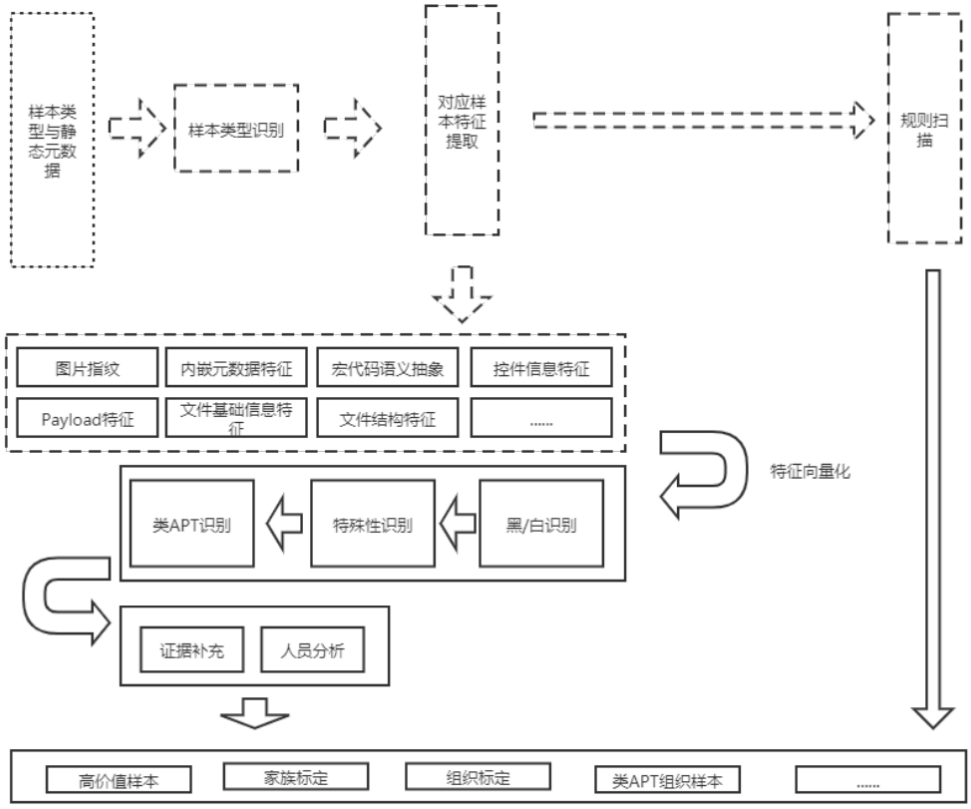

图 2 文档类样本处理流程图

其中特征抽取过程抽取了包含文档内图片的指纹、类型，以及对应的处理器指纹等，同时包括文档的内嵌数据：OLE 对象、控件对象等；如果含有宏代码，会对宏代码进行进一步的语义分析与代码抽象；如果文档含有漏洞，会抽取漏洞部分的二进制数据等。基于这些抽取出来的数据，共提取近百种特征，这些特征会在后续的黑白识别、特殊性识别与类 APT 识别中分别投入到不同的机器学习模型进行分析判断。

通过模型判断出来的样本在经过后续的证据补充和人工分析环节完成对其真正的定性判断，最后生产出高价值样本/类 APT 组织样本以及标定了组织、类别、家族的样本。

其中采用的数据都是经人工审核、厂商报告提取、多种杀软交叉验证得到的数据集，经过图像模糊度计算相关算法、语义抽象相关算法等多种算法来进行特征提取，通过提取后的特征结合安全研究员在日常分析中针对不同阶段和不同类型的样本侧重点方面的经验来进行分模型训练，然后按顺序进行多模型组合判断，从而达到层层过滤的效果。

【应用效果】

本系统架设在公司内部，每天处理近 50 万样本，能从中快速识别并推荐已知 APT 攻击组织的样本并且在一定程度上识别一些使用新技术的攻击样本，同时能快速标定一些黑产及僵木蠕对应的组织及家族。

【下一步工作建议】

在对抗一些新的利用方式上需要不断完善，对特征提取和文件静态信息提取的颗粒度需要进一步提升。此外，机器学习在鉴别黑白与在黑样本中寻找类 APT 样本或者特殊样本方面同样存在应用空间。同时为了弥补单一同源哈希造成的漏报现象，在后续会引入特征组的方式进行多特征相似计算，来进一步归因。

基于威胁情报的多维恶意域名自学习检测技术

【场景描述】

黑客攻击日益呈现产业化、规模化、自动化的趋势，对百姓财产和公共关键基础设施造成的危害正持续增长。然而传统的恶意域名检测算法主要通过纯域名的静态特征检测，缺少自学习能力、选择的动态特征维度不够全面等问题，对于复杂的域名流量，几乎不可能完成有效的检测识别。

针对传统方法在特性选取不足、无自学习能力等问题，绿盟采用一些新型的智能技术，基于机器学习技术对海量威胁情报数据进行智能分析，从繁杂的海量信息中提取出高价值威胁特征，并且对看似不相关的多方位情报进行横向、纵向关联，深度挖掘多维线索之间隐藏的内在联系，进而对系统的整体威胁态势进行清晰的描述。

【技术方案】

本项目拟采用基于多维度融合分析的恶意域名智能检测方法解决传统方法难以解决的问题，技术框架图如下：

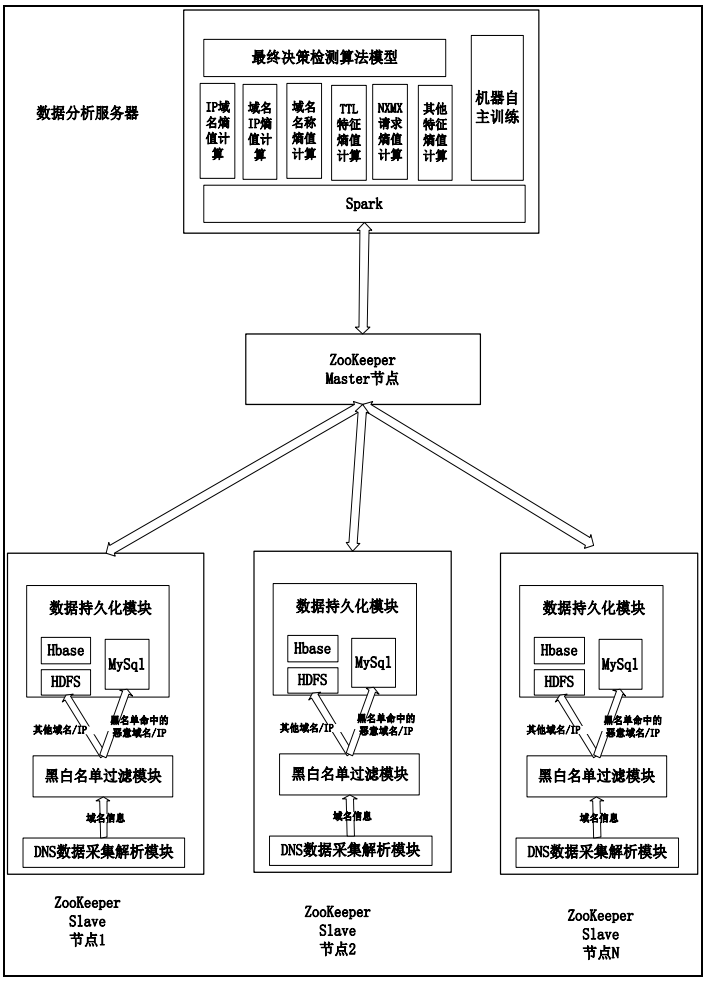

图 1 多维恶意域名检测技术框架图

利用大数据处理技术，建立基于六个维度特征对恶意域名进行静态、动态全特征检测的算法模型，提高恶意域名识别的效率与准确率。从域名名称、域名对应多个 IP、IP 对应多个域名、TTL 特征、MX/NX 请求特征以及其他特征（特殊 IP/子域空间）等六大维度提取特征建立恶意域名检测模型，实现对恶意域名的全特征综合检测，可解决现有方法中的对恶意域名检测的维度单一导致检测的召回率、准确率相对较低的问题。

同时建立恶意域名特征自学习的智能算法模型，使系统同时具有自学习能力，根据黑白训练样本集自主学习恶意域名特征，减少人工干预的工作量，提高生产效率。在 n-gram 机器学习算法的基础上加以改进建立新的恶意域名特征自主学习算法模型，检测过程全程自动化，无须人工介入设置参数。自学习的多维恶意域名检测技术实现检测过程全自动化，有效提升识别效率和准确性，能够从海量流量和日志中生产恶意域名情报。

【应用效果】

基于威胁情报的多维恶意域名自学习检测技术有效提升恶意域名检测的准确率和时效性，及时在海量数据中发现恶意域名避免用户对业务感知产生负面影响。

【下一步工作建议】

绿盟威胁情报已普遍应用于各种产品中，取得了良好的效益。但仍存在不足，需要持续改进。后续工作将聚焦于如下方面：扩展互联网层面的情报采集，进一步丰富情报内容；进一步研究更高级的情报分析和理解算法、情报数据隐含的威胁行为和发展态势分析技术；推动情报共享标准和技术规范的制定，构建绿盟与企业用户的情报共享平台和网络。

3.1.8 态势感知

基于人工智能的恶意软件攻击态势感知系统

【场景描述】

网络安全的态势感知是攻防对抗的重要关键手段，除了基于对威胁的准确识别，还需针对威胁与攻击进行趋势和影响分析，达到预测的进一步高阶目标。攻击者为避免被检测、阻断，常采用新型或者变种的恶意软件，利用 0Day 漏洞发起攻击。传统基于特征的检测手段从有效性、时效性角度来看都不足以进行对抗。

针对新型攻击难以识别、判定的问题，中兴通讯创新性的研究出基于人工智能的恶意软件判定方法，可以对网络中传播的恶意软件进行有效识别，快速提取攻击特征，并结合资产属性与攻击者信息进行威胁程度和影响态势预测，为安全设备进行防护提供重要依据，保障 IT、OT 等各种信息化网络安全。

【技术方案】

一、技术方案概述

方案总体由恶意软件检测和攻击态势预测与呈现两部分构成，为安全运维人员提供准确的威胁判断和风险感知。架构如下图所示：

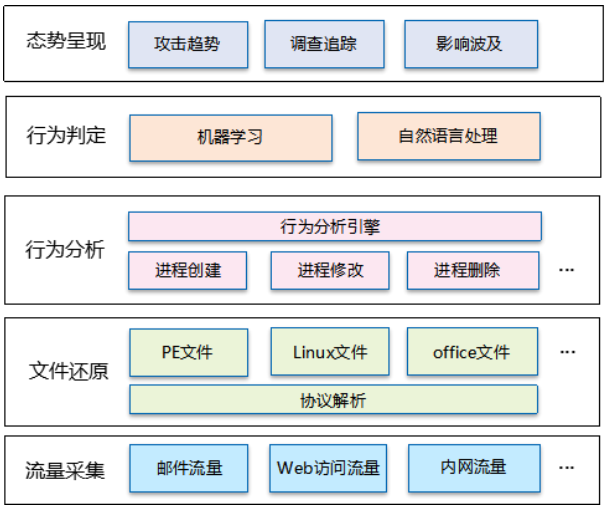

图 1 恶意软件攻击态势感知整体架构

整体方案包括如下几个部分：

- 流量采集：系统探针全方位采集监测网络的进出口及内部流量，包括邮件流量、Web 方位流量及其它内部传递的流量数据；
- 文件还原：系统探针同时进行流量协议解析，还原流量中传播的各种类型文件；
- 行为分析：还原后的文件被传递到行为分析引擎中，引擎在虚拟机环境下触发文件运行并捕获运行过程中的各种操作行为；
- 行为判定：人工智能检测引擎依据训练好的判定模型，对文件的一系列行为进行判定，确定文件是否为恶意软件；
- 态势呈现：系统对网络内爆发的 TopN 恶意软件及爆发趋势进行预测与呈现，同时支持对具体恶意软件来源及攻击目标的调查追踪。结合网络内的资产属性和威胁情报对恶意软件影响范围进行预测与直观性感知。

其中，行为判定模型训练及检测工作流程如下图所示：

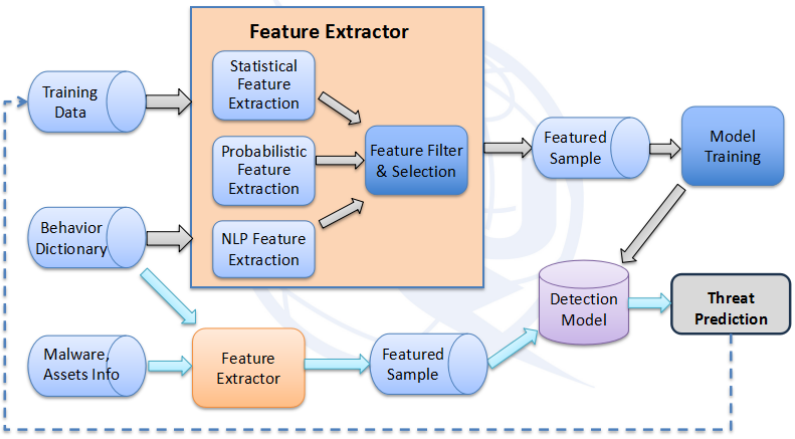

图 2 模型训练及检测工作流程

整体工作过程包括两个智能化部分：

- 特征提取：使用自然语言处理的方法提取隐含的关系型行为特征（例如：tf-idf、isi、ida、nmf 特征等），使得描述文件、资产和攻击的维度更丰富。
- 分类与预测方法：可采用二分类模型来快速确定文件是否为恶意进行文件样本分类，使用随机森林（Random forest）、增压模型 xgboost 等机器学习算法来进行资产遭受攻击可能性概率的预测。

二、技术方案优势

本方案的优势在于综合提取样本的统计特征、概率特征及 NLP 特征并进行样本描述，丰富的特征使得训练出的判定模型准确度更高。

本方案克服了传统基于规则的恶意软件检测与应对方法的以下问题：

- 专家经验的缺失导致绝大部分软件的性质无法确定；
- 基于固定某个特征或者特征组合的判定方法误判率比较高；
- 特征提取需人工参与，规则生成周期长；
- 无法对可能被攻陷的资产进行态势感知，无法指导安全管理人员进行优先级排序与处置；

方案在提供准确文件恶意性判定的同时，针对应用环境与资产属性，结合威胁情报丰富的攻击方特征对检测出的恶意软件进行非实用全面的态势呈现，提供可能实际遭受攻击的资产影响范围、爆发趋势态势预测，如图 3 所示。

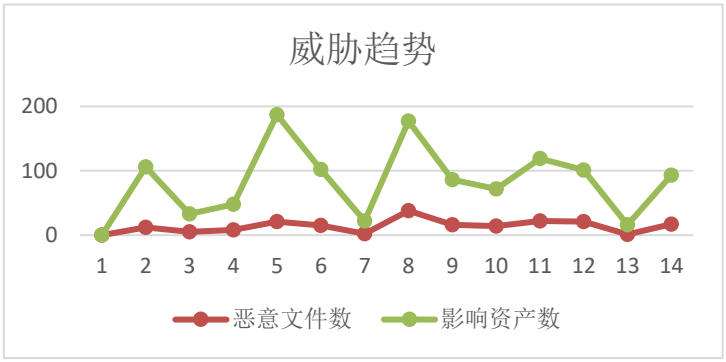

图 3 恶意软件爆发与影响趋势图

对于某个具体的恶意软件，方案提供诊断工具确定恶意软件的攻击过程及波及范围，预警最可能遭受攻击的资产以及核心业务高危资产与人员，从微观角度对具体某个恶意软件攻击过程及进展态势进行直观感知。例如图 4 中展示的样例，为外部 163 邮箱携带并投放的命名为产品相关介绍 word 附件，检出该附件为恶意文件后，可获知所有被投放方信息，并判断可能被攻击成功的重点高危资产，向安全管理人员进行告警与整体态势呈现。

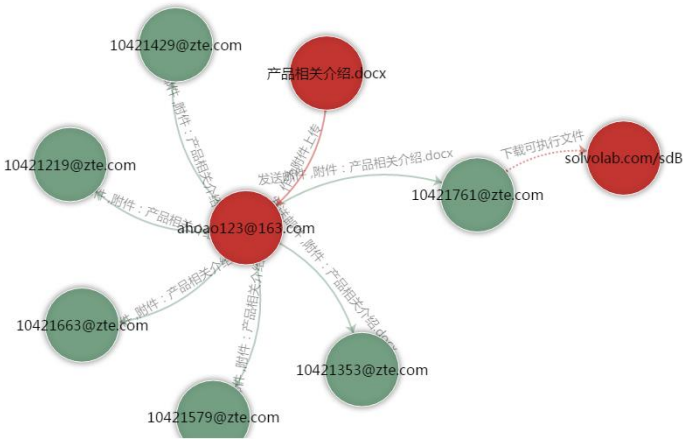

图 4 恶意软件攻击态势呈现

【应用效果】

本方案自 2016 年起，在金融、电信、政府机构等选取试点启动上线应用，应用期间在不断提升恶意软件判别准确率的前提下，进行应用场景下的态势预测，应用期间协助数百起重大网络安全事件的发现与响应。例如：2016 年全球勒索软件爆发年，第一时间发现全球最新勒索软件事件及其变种，预测影响范围，预警应用单位保障了信息安全，避免了大规模攻击后果；发现针对世界 500 强企业高管的定向攻击事件；发现针对电信企业地域性定向攻击事件；进行内外部追溯与态势分析，协助 WannaCry 大范围攻击事件的应急响应，并持续发现 WannaCry 最新变种；发现 2019 勒索病毒 GandCrab5.2 最新变种。本方案极大提升了对未知威胁的检测预警与响应能力。

【下一步工作建议】

不断积累有效训练样本，对算法进行调优：攻击和环境都在不断演进，固化的训练样本会导致模型的适配性不好、判别准确率下降。需要人工参与，不断对样本进行筛选、标记，持续优化训练集。尤其在预测方面，仍需不断探测新的算法模型，随着训练数据和特征的量级提升，考虑采用深度学习的推荐系统模型来进一步尝试。

智能大数据分析及态势感知

【场景描述】

企业面临的安全威胁方式和数量正在发生巨大的变化，传统的安全防护系统（包括防火墙、入侵检测、防病毒、抗 DDOS、WAF、漏扫等）+安全分析平台（包括 SIMS、SOC、审计分析、合规分析等）的安全管理模式面临严峻挑战。

Ultra-SecSight 是神州泰岳安全公司发布的企业信息安全大数据分析及态势感知平台，基于“数据驱动安全”的全新技术理念设计，通过对于多种异构的安全数据集中化融合、汇聚，并利用大数据相关的先进技术，通过科学的数据挖掘、智能分析、大数据算法的运用、实时/离线分析处理、机器学习和数据可视化等技术手段实现信息安全大数据态势感知能力的建设。

【技术方案】

一、现状描述

各个安全平台、安全设备独立输出、管理的安全数据，包括：事件数据、日志数据、基础数据、威胁情报数据等，由于独立建设、分析、输出及管理，安全风险分析呈现了分散化、碎片化的特点，导致当前安全分析准确性和可信度较低，严重影响安全风险处置及管控工作效率。

- 安全数据来源多、数据结构复杂，难以通过统一维度视角进行分析，缺乏自动化数据融合手段，整理为集成度、价值性较高融合安全数据；
- 安全风险分析结果还存在误报率偏高、可信度较低，缺乏灵活交互、智能分析手段，以支撑安全分析决策；
- 安全防护能力依赖安全防护规则或特征库，而防护规则或特征库的更新滞后于攻击手段的变化；
- 针对未知的安全威胁，只能在安全事件发生之后进行亡羊补牢式的被动响应；
- 安全系统之间缺乏有效整合和联动机制，安全防护范围存在盲点。

针对以上，泰岳 Ultra-SecSight 态势感知平台提供将多种、更灵活的安全风险数据汇聚、处理、分析能力，输出更精准、更具价值的风险告警、感知数据，充分体现了多源、异构、丰富风险数据的综合、关联分析需求。

二、多源异构安全数据融合

Ultra-SecSight 全面采集企业 IT 管理支撑相关的人员、资产、事件、日志、配置、策略、流量等数据，最大程度的将企业信息安全管理工作通过数据实现可视化，支撑上层风险分析需求。

平台基于汇聚多源数据的优势，基于自动化异构数据融合手段，针对某大类安全数据，通过对多源、异构安全数据进行数据融合，解决多来源数据冲突、属性分散，单来源数据不完整、片面等问题，并针对数据实现过滤筛选、补全、转换、聚合归并、解析抽取等数据处理过程，综合构建基础的安全数据中心。

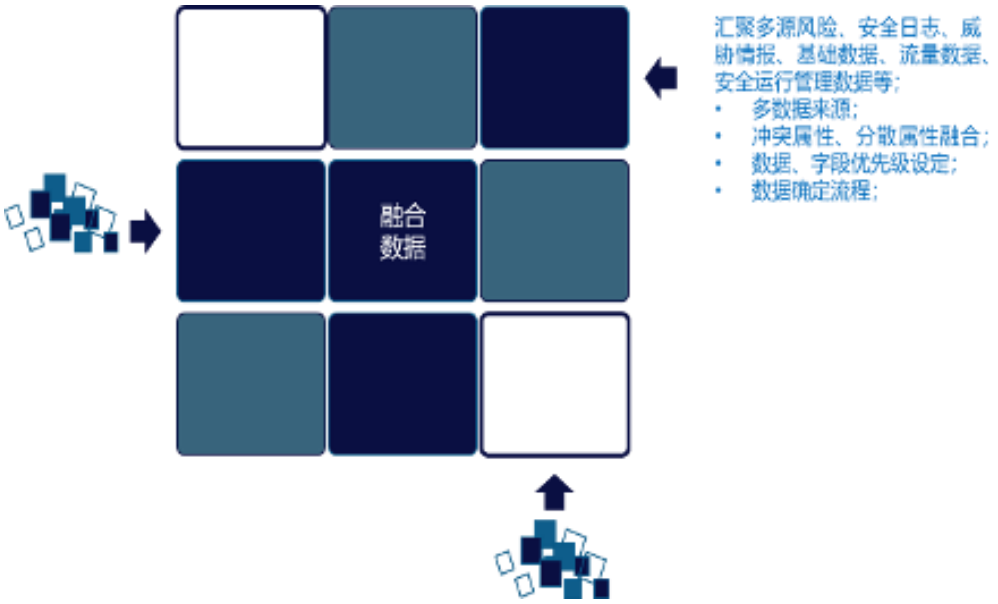

图 1 多源异构安全数据融合架构图

由于安全风险最终会落到资产上，所以帮助企业摸清资产数据是态势风险分析的基础，企业分散建立了各种资产管理系统，但不同资管系统由于人工维护、管理侧重点不同、业务目标不同等原因，通过单一来源获取资产数据往往会导致数据不可用、不完整或错误等情况，所以 SecSight 基于汇聚多源数据的优势，通过数据融合手段，构建 IT 基础资产数据中心。

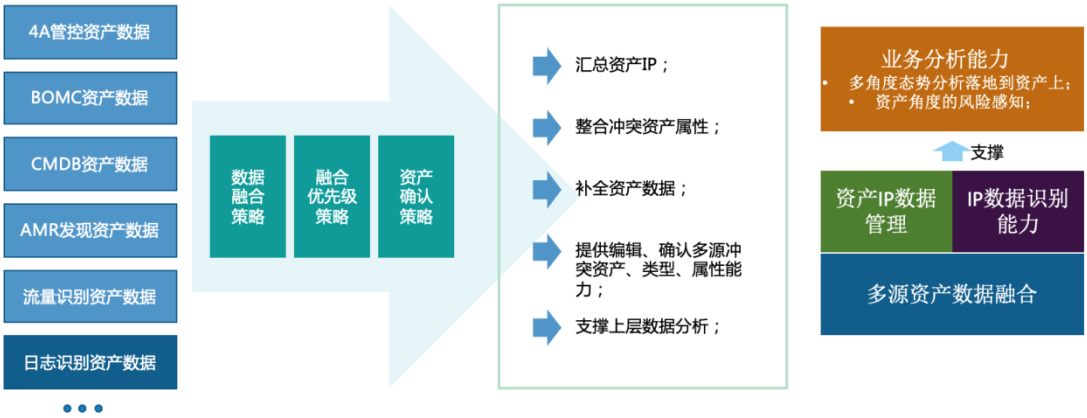

图 2 多源资产数据融合

三、安全告警智能分析

基于融合企业安全数据，在汇聚各类安全系统、安全设备、安全检测工具等输出安全事件的基础上，通过深度理解、挖掘各类安全数据间相关度、关联方式、逻辑关系，帮助提升安全风险告警输出的可信度，站在企业整体安全视角综合分析安全事件数据，实时输出更精准、更可信的安全告警，并基于基础 IT 数据，快速定位风险资产、责任人。底层支持面向安全风险告警分析的智能模型构建，智能模型包括针对时间序列的风险分析模型以及针对风险的黑/白风险指纹库的训练及过滤模型。

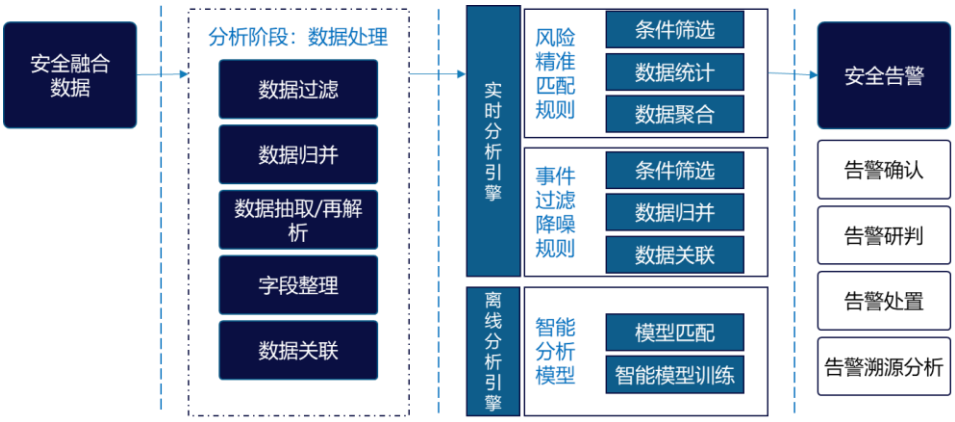

图 3 安全告警智能分析架构图

由于安全风险管控最重要的核心绕不开企业资产，所以基于资产视角的安全威胁分析态势是 SecSight 核心的分析能力及专题分析场景之一，结合泰岳多种资产发现、检测手段，输出资产安全分析能力。通过对内外部威胁情报进行有效整合，持续性提升安全感知与处置能力；基于威胁信息的模型自主更新能力，构建安全风险管控的持续改进能力。

四、安全风险闭环处置

移动以态势感知作为安全分析与管控的入口，针对输出的安全风险告警，构建了由多个安全风险小组，包括安全监控组、专家研判组、风险处置组等组成的安全风险处置架构，针对不同风险类型、风险级别，支持多种一键处置能力，包括：处置工单派发、快速处置联动、安全设备处置联动等，建立了安全风险处置闭环。

图 4 风险处置流程图

其中一键应急处置是帮助企业根据安全运维现状，结合安全威胁集中处置思路，建立安全告警日常风险处置、安全保障阶段的串联的快速的安全响应能力，实现安全风险的闭环管理。一键应急处置通过转变安全应急处置方式，提升安全应急处置效率，通过统一管理安全告警处置任务，集中安全风险处置入口，从而简化了应急处理流程，减少了应急处置人员，是集中安全管理的有效实践。

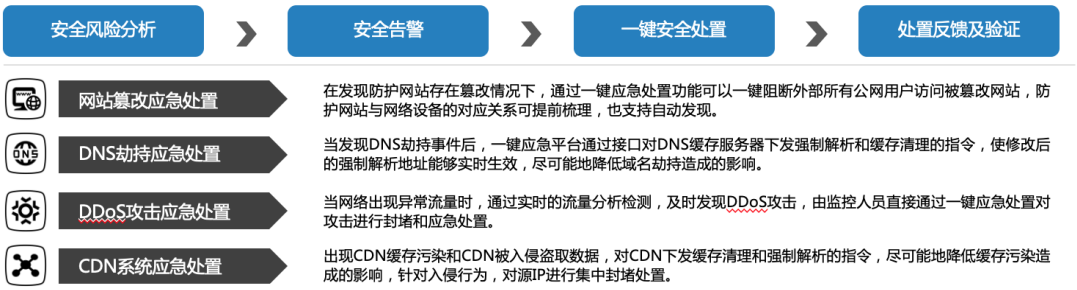

图 5 一键应急处置场景及流程

【应用效果】

基于汇聚多种各类型风险数据、IT 基础数据，经过安全告警智能分析，在风险可视化阶段，支持风险告警数据、风险告警可视化视图及场景的输出，将分析过程快速、高效的转化为面向安全分析用户、数据分析用户、安全运维用户、各业务管理用户、领导等多种角色用户的风险告警展现输出、可视化感知能力，帮助用户快速定位风险点，串联后续风险闭环处置流程。

【下一步工作建议】

持续丰富风险分析能力，并构建多维分析视角，提供交互式灵活分析功能，输出多视角的业务感知场景，全面支撑企业安全决策！

同时针对 5G 时代，不断支持多种 5G 协议、数据的扩展，扩展安全风险分析模型及应用场景，紧跟时代步伐。

多智能分析引擎的态势感知

【场景描述】

随着企业信息化资产数量日趋增多、系统的关联性和复杂度不断增强，信息安全防护工作面临前所未有的困难和挑战。目前传统态势感知设备仍存在分散设备难以统一管理、海量日志分析困难，判断决策缺乏依据、事件响应延迟等问题。

针对以上问题，绿盟发挥在态势感知领域多年的创新优势与技术和积累，研发了一套全新的态势感知系统。系统利用大数据技术结合威胁情报进行集中处理、关联分析，再利用可视化技术，将各种安全事件进行可视化呈现，为安全运营提供可靠的信息数据支撑。

【技术方案】

一、系统架构

态势感知系统逻辑架构图如下：

图 1 系统逻辑架构图

1、 最底层是大数据技术平台，为态势感知系统业务提供运行环境及各种组件和资源的管理框架；

2、 能力组件层，为业务功能提供能力特性支持，包含行为与威胁分析子系统、资产运维子系统、脆弱性管理子系统等；

3、 最上层是业务功能层，主要包括全景态势、运维响应、威胁管理、脆弱性管理、环境感知、威胁情报、合规审计等内容。

在上述态势感知系统中通过各类安全设备或探针，结合集中管理平台，为了提供更为智能的应急响应和安全服务，引入多种智能分析引擎举例如下：

- NetFlow 机器学习引擎：经由分析 Netflow 收集到的资讯，网络管理人员可以知道封包的来源及目的地，网络服务的种类，以及造成网络壅塞的原因。绿盟安全分析研究人员将 NetFlow 数据用作安全类溯源和机器学习研究，从而实现流量拼接、密度聚类、爆破检测、蠕虫检测等。
- 机器学习驱动的数据分析：通过行为特征搜集、素材准备、行为特征量化、特征向量选择、分类、聚类、回归算法应用以及模型验证和调优等过程，实现海量数据的分析、挖掘；
- 行为与威胁分析子系统：在威胁建模知识库基础上，实现实体行为的分析及威胁分析和推理，完成从日志到事件乃至于具体威胁场景或攻击过程的转化输出。

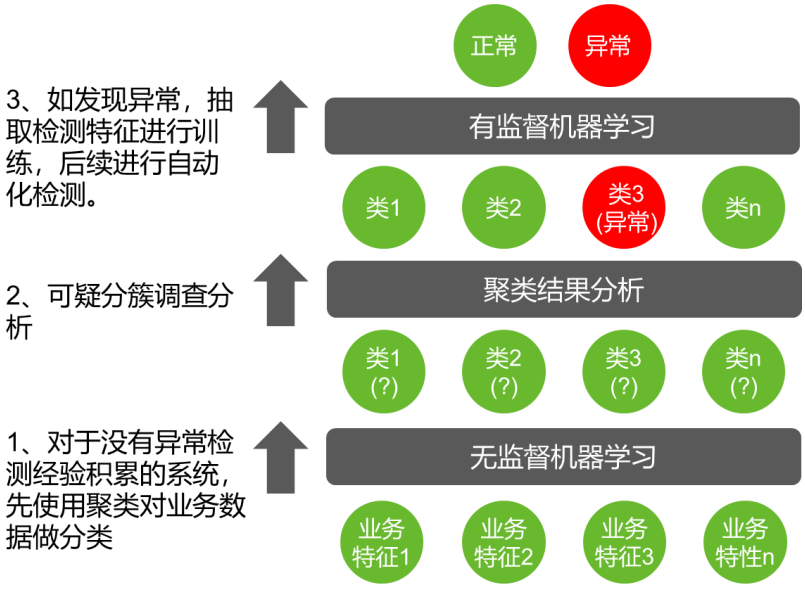

图 2 基于机器学习技术的行为分析

围绕安全攻防本质，聚焦安全攻防场景、机器学习算法等实战性的业务需求来构建未来网络安全防御体系，提升针对网络安全态势感知、监测响应处置能力，实现安全态势全方位感知。

【应用效果】

依托于态势感知平台建设，在部分运营商通过分光链路带宽监控 700G+业务流量，已经形成了威胁攻击溯源、全流量数据分析、一键封堵处置等实战化安全分析及响应能力，在重要活动保障期间已经成为安全运维的主要手段，并实现多次 0 重大安全事故。

【下一步工作建议】

下一步的计划主要包含如下方面：充分利用网络全流量、大数据分析及机器学习技术，大幅提高安全事件监测预警和快速响应能力；更加聚焦未知安全威胁，不断提升针对高级持续性威胁事件的态势感知。

3.2 内容安全篇

3.2.1 骚扰诈骗电话检测

骚扰电话机器人自动应答技术应用与实践探索

【案例简介】

作为一家基础运营商企业，中国移动长期开展骚扰电话治理工作，针对近年来群呼类骚扰电话数量大、智能化的趋势，创新推出高频骚扰电话防护服务，并研发机器人自动应答技术，实现骚扰电话机器人代接，

以机器应答对抗机器群呼。一方面，通过柔性挂断和内容提醒有效提升用户感知，避免误拦通话对用户造成的影响；另一方面，音频可进一步提升骚扰电话治理的精准性，促进提升产品竞争力，增强用户黏性。

“骚扰电话人自动应答应用”实践是人工智能赋能各行各业形势下的一种创新安全服务，通过智能网拦截代接技术，实现人工智能应答技术在骚扰电话防护领域的应用，有效保障广大用户的通信权益，在整体方案、关键应答技术等方面实现重大突破，为全球运营商和互联网公司分享一种人工智能安全治理思路，合力净化通信网络环境。

【场景描述】

近年来，骚扰营销电话层出不穷，严重影响广大人民群众的日常生活。不法分子受利益驱使，花样翻新，采用机器人群呼等新型手段传播骚扰电话，成本低、频次高、数量大且研判困难。为履行企业社会责任，保障广大用户通信权益，中国移动持续开展骚扰电话治理工作。针对骚扰电话群呼新形势和千人千面的特点，安全治理团队颠覆传统治理模式，创新提出将骚扰电话的拦截权交还给用户的理念，及时地在业界率先推出“高频骚扰电话防护服务”。利用网络侧拦截技术，结合用户个性化设置，对高频骚扰电话、商业营销骚扰电话等进行实时拦截，保障用户合法权益。2019 年 3 月推出以来，该项服务累计为 1488 万用户拦截骚扰电话 27 亿次，用户反映良好。

但由于骚扰电话的复杂性，在高频骚扰防护服务过程中，硬性拦截通话难免存在感知局限和误拦风险。与此同时，人工智能技术发展迅速，中国移动九天人工智能平台在自然语言处理和智能语音方面已有多年技术积累，输出智能客服系统、会议语音转写系统等多项产品，均具有良好的使用效果。综上，针对骚扰营销电话这一行业难题，在高频骚扰电话防护中引入人工智能应答技术，用机器对抗机器乃大势所趋，骚扰电话治理已进入人工智能对抗时代。

【技术方案】

一、中国移动人工智能应答应用思路

基于上述背景，中国移动迅速展开机器人自动应答研究开发，计划在高频骚扰电话防护的基础上，提供给用户一个可代接高频呼叫的人工智能助手。用户启用后，人工智能助手将与骚扰营销主叫开展数轮智能对话，实现黑名单通话柔性拒绝，同时，用户可以在微信公众号中查看对应的来电信息及通话语音内容。通过该技术应用，一方面，可有效提升高频骚扰电话防护服务的用户感知，避免误拦通话对用户造成的影响；另一方面，音频可进一步提升骚扰电话治理的精准性，进一步提升产品竞争力，增强用户黏性。

二、人工智能应答实践探索内容

1、应用方案和网络架构

如图 1，通过在现有高频骚扰电话防护平台引入媒体处理模块，实现机器人自动应答。对于未启用人工智能助手的用户，高频防护服务于原有流程保持一致；对于启用人工智能助手的用户，由机器人代接命中黑名单的呼叫，与来电方进行数轮简单对话后结束通话。

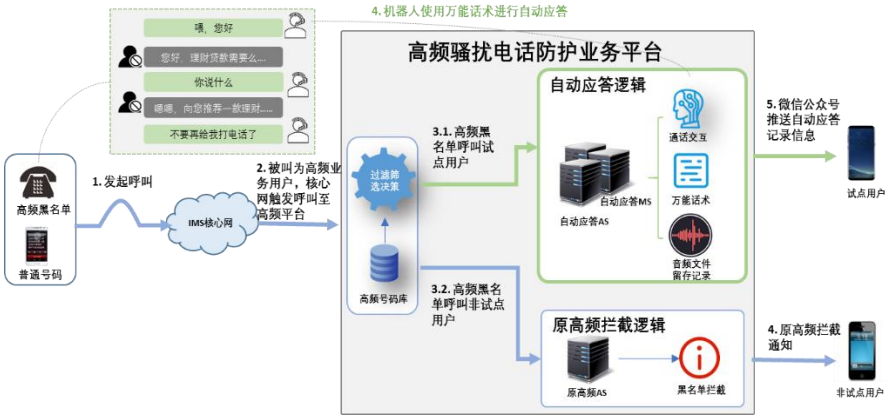

图 1 机器人自动应答应用方案

网络架构方面，在原高频骚扰电话防护平台基础上，新增 MS 媒体服务器，并引入 VAD、ASR、NLP 等技术模块，针对广告营销类电话，由平台将呼叫路由至人工智能机器人自动应答；针对其他正常用户发起的呼叫，则正常放通。

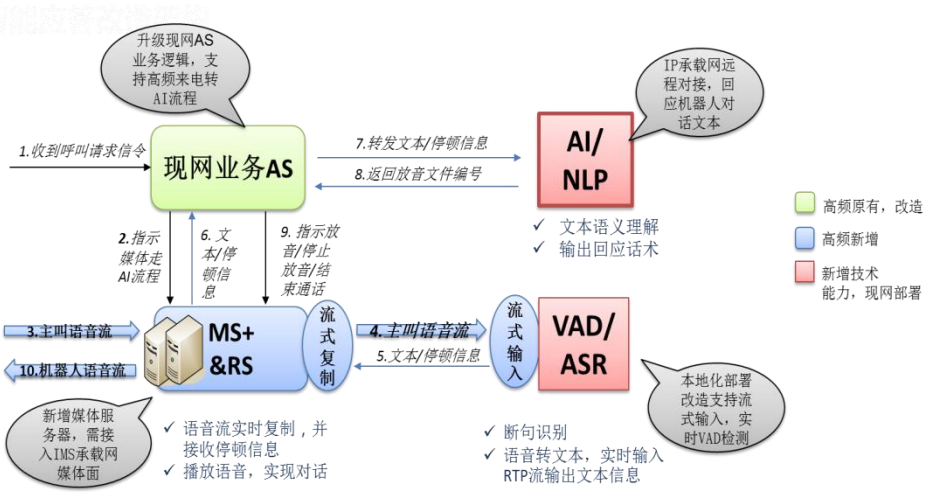

图 2 机器人自动应答网络架构

具体地，网络架构方案主要包含四大步骤:

- 呼叫控制模块收到呼叫请求信令，指示媒体处理模块走人工智能流程；
- 媒体服务器将语音流实时复制到语音识别模块；
- 语音识别模块将语音转为文本形式，并识别停顿信息；
- NLP 模块接收文本及停顿信息，进行标签分类及语义理解，并返回话术编号及交互状态标识。

2、关键技术探索及前期实验

机器人自动应答技术上线前，专家团队依托中国移动九天人工智能平台开展了充分的技术研究和测试。“九天”人工智能平台赋能安全中台，助力构筑中台的安全云脑引擎，面向防骚扰机器人等内容安全管控场景，从能力引擎到创新应用开展深度合作，护航“5G+AICDE”安全。本次主要涉及机器人自动应答场景识别、语音识别、对话管理、断句识别等关键技术。

图 3 九天人工智能平台实验内容

经充分验证，技术初步具备应用条件。进而，项目团队结合骚扰电话对话特征，对关键技术进行了多轮改造优化。作为自动应答中的关键环节，NLP 模块承担语义理解、多轮对话管理等重要功能，是整个防骚扰电话机器人的核心。其管道式架构如下图所示：

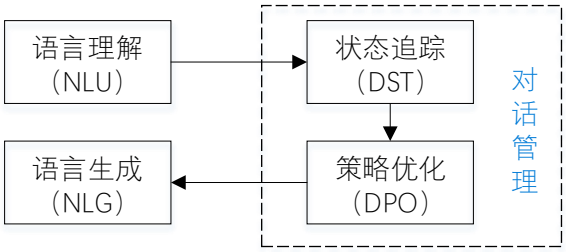

图 4NLP 管道式处理架构

该技术模型包括四个子模块:

1) 语言理解主要识别不同层级用户意图，包括一级标签（骚扰/日常事务等类别标签）、二级标签（骚扰-房产销售/骚扰-理财推荐等）。系统使用多模式匹配和基于深度学习的文本分类相结合的算法对文本输入进行多级标签分类。
2) 多轮对话管理包括对话状态追踪和系统动作决策，状态追踪在对话的每一轮中，根据对话历史得到当前对话信息的状态表示，防骚扰电话机器人主要采用槽位追踪的方式捕获对话状态，如<位置，北门>等。
3) 策略优化旨在促进系统选择正确的系统动作，基于对话意图、对话状态追踪结果和当前策略选择系统动作，在交互过程中，根据外界返回的奖励值不断进行策略优化。
4) 语言生成主要根据系统动作返回对应的系统回复话术，传递到后续语音合成模块。

3、用户端功能设计

作为一款面向用户的服务产品，项目团队基于微信公众号进行了功能配套开发，以交互友好为界面功能设计目标，对试点用户的界面操作和代接设置进行开发和优化。

图 5 自动应答服务界面

【应用效果】

机器人自动应答项目于 2019 年 11 月启动试点流程，2020 年 3 月在现网升级上线，自动应答内部试点用户的通话实现机器人代接。自动应答音频平均时长 25 秒，其中，系统黑名单通话代接占比 89.5%，号段拦截代接占比 7.5%，个人黑名单拦截代接占比 3.0%。经数据分析和用户回访，试点情况整体良好。

该项技术的应用，在原有高频骚扰电话防护的基础上，实现了营销电话柔性拒绝，并为用户提供拦截录音，有效提升用户感知，将进一步提升用户黏性，助力品牌重塑。同时，应答音频一方面积累了高精度的号码标记库，另一方面促进了骚扰电话治理闭环处置，对于误拦的快递送餐类号码，一经发现立即出库；对于违法、涉黄等号码，本网号码转交由骚扰电话集中管控平台进行关停、加黑等源头处置。闭环治理如图 6。

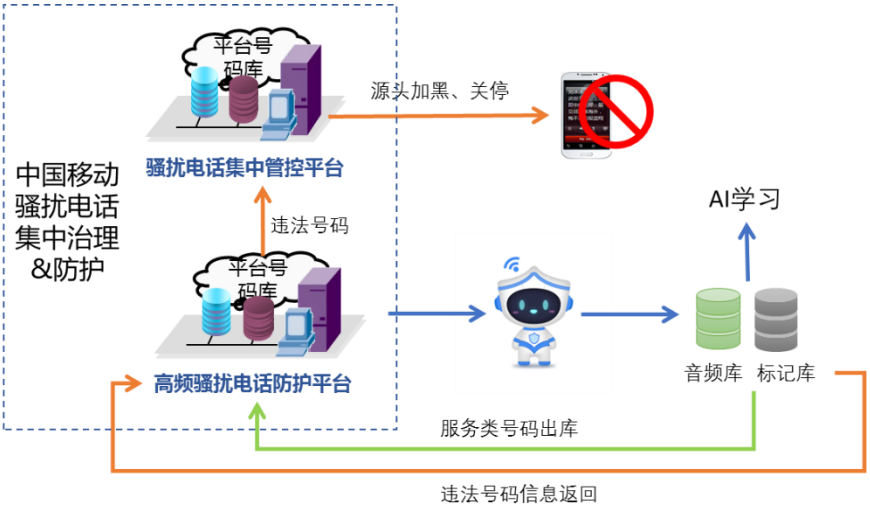

图 6　自动应答助力实现闭环治理

【下一步工作建议】

“骚扰电话人自动应答应用”通过智能网拦截代接技术，实现人工智能应答技术在骚扰电话防护领域的应用，在整体方案、关键应答技术等方面实现重大突破，有效保障广大用户的通信权益。该项实践基于运营商通信网络，使用人工智能应答技术，并提供给用户互联网产品的使用体验，是一款“人工智能+通信网+互联网”的安全防护升级服务，为骚扰电话治理和运营商营销号码分类都做出很大贡献，为全球运营商和互联网公司分享一种人工智能安全治理思路，合力净化通信网络环境。

人工智能反欺诈系统

【场景描述】

传统骚扰、欺诈识别模型的设计思路是通过研究电信欺诈模式形成欺诈剧本，将诈骗剧本翻译成通话行为规律，然后对通话详单的各个行为指标设置异常阈值，形成较为固定的识别策略，不断优化。这种方法存在以下不足：

1. 依靠对欺诈剧本的主观理解；
2. 设置的阈值较为粗糙；
3. 可初步识别异常行为；
4. 调优过程复杂。

同时随着不法分子利用人工智能技术拨打“机器人骚扰电话”以及新型诈骗，欺诈剧本愈发复杂，不断绕过固定的识别策略，给传统通信欺诈识别方案提出挑战。基于人工智能的反欺诈的设计思路，是将前期已识别的号码设为种子，然后对全量数据与种子数据进行自然学习，最后输出模型阈值反向优化。这种方法存在以下优点：

1. 对种子数据行为特征的自然学习，更为准确；
2. 多层深度学习，阈值精准；

针对以上问题，中国联通发挥创新、研发优势，研发基于分布式计算的智能信息通信欺诈治理系统，在各分子公司实现产业落地。同时响应国家和社会要求，一方面，提供公益性、基础性服务，全面覆盖个人用户；另一方面，开放合作，向行业用户有偿提供防欺诈服务。

【技术方案】

中国联通搭建了基于 Acumos AI 的人工智能反欺诈模型孵化实验平台。该平台可为开发人员提供方便快捷的开发工具、可视化与模块化的开发环境、开发模型一键部署功能，有效降低开发门槛，提高开发速度。本平台整体系统架构如下图所示：

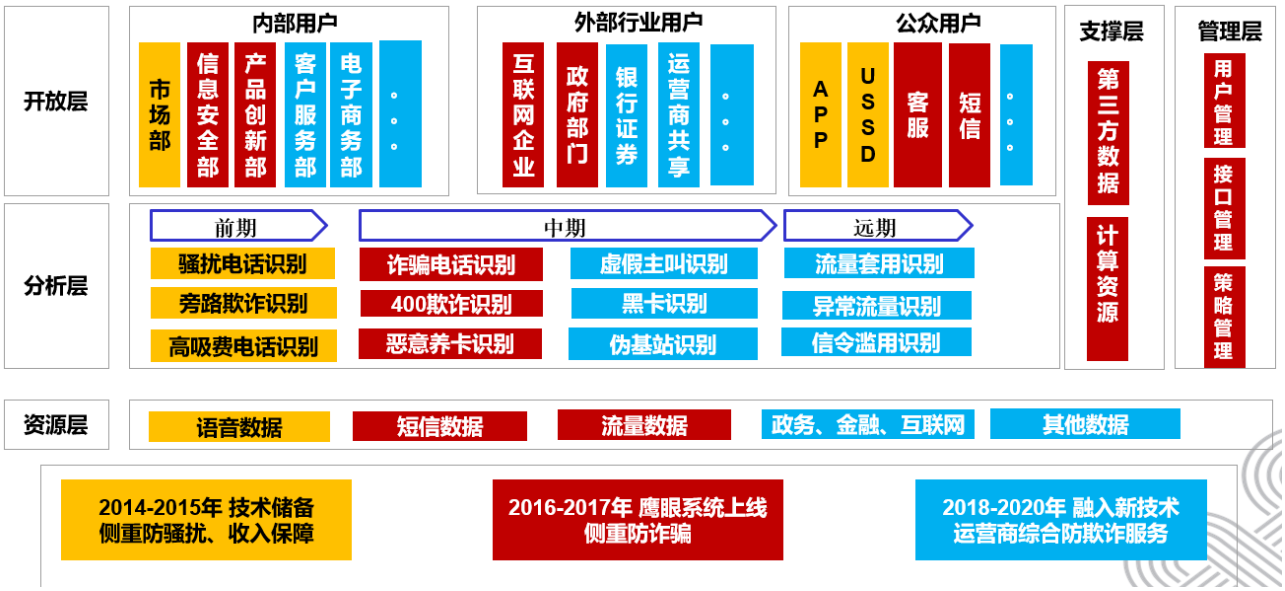

图 1 人工智能反欺诈系统架构图

路线：对 B 域的语音、短信、上网记录话单以及合作方的交换数据进行分析建模，孵化各类欺诈识别对模型；将模型输入现网数据中心，建立信息通信欺诈号码库；通过短信、电话、客户端、USSD 等多种方式对用户进行有效提醒，减少用户经济损失。

- **人工智能+电信防欺诈整体研发思路：**

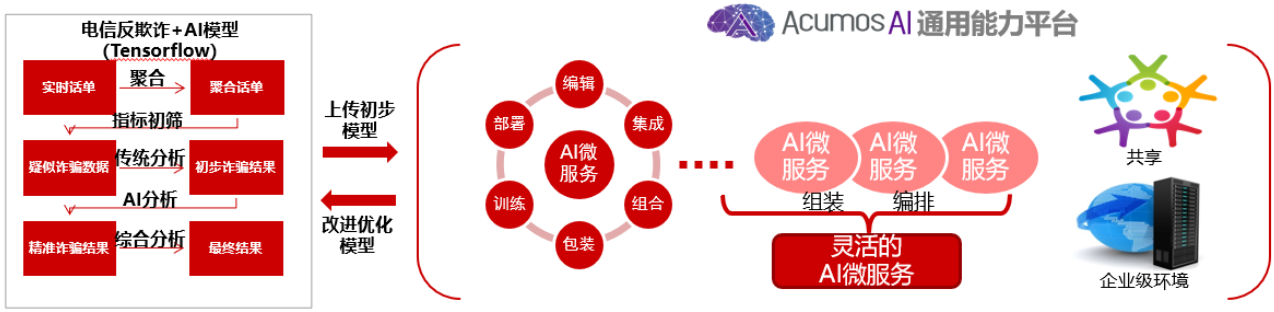

图 2 人工智能反欺诈系统研发思路

根据养卡的语音、短信、流量三域通信行为特征，利用自有数据及整合合作伙伴数据，应用大数据分析技术，识别养卡号码。具体步骤如下：

1. 利用前期集团防欺诈系统研发建设过程中积累的数据成果作为种子，利用人工智能深度学习算法进行自学习，形成人工智能智能识别模型；

2. 将模型输入至人工智能微服务平台（基于 Acumos），形成人工智能微服务的一整套闭环能力，实时识别欺诈行为、不断自学习优化模型。

- **人工智能+电信防欺诈研发路线：**

1. 内部孵化

利用人工智能+前期防欺诈成果，初步实现人工智能欺诈识别模型，并提高原有机器学习技术的识别率。

2. 现网验证

接入现网环境，进一步优化反欺诈人工智能模型，实现基于实时话单欺诈行为分析，经过自学习过程，不断优化改善成熟模型提高识别率与覆盖率。

3. 产品化

通过结合开源工具与自主研发，基于电信反欺诈模型，打造成熟人工智能通用能力产品。

【应用效果】

实现效果：目前已开发两套新型人工智能反欺诈模型，分别为“基于标定数据的 DNN 全连接神经网络模型”与“基于实时数据的 RNN 模型”。可实时优化、动态调整模型并进行对抗识别，可自学习参数，识别准确率会随训练量增长逐步提升。

经实验验证，从传统机器学习方式的 70%的预测率提升至 88%左右，并根据多层算法不断优化准确率至 95%以上，新型人工智能反欺诈识别模型可自适应变化多端的欺诈剧本，从而满足反欺诈生产环境要求。

1. 对江西、四川、宽带冠币、宽带移动支付等恶意养卡重点地区与用户群进行分析，共发现疑似恶意养卡号码 4G 用户 10 万个、2/3G 用户 13 万个。

2. 对公司 400 号码话单进行分析，得到疑似异常 400 骚扰号码共 5 万个。

【下一步工作建议】

1. 基于现有的人工智能欺诈识别模型，接入现网环境，进一步优化反欺诈人工智能模型，实现基于实时话单欺诈行为分析，经过自学习过程，不断优化改善成熟模型提高识别率与覆盖率。并且通过结合开源工具加强自主研发力量，结合深度学习、图数据库、卷积/图卷积神经网络等新兴技术研发新型识别模型，加强新型诈骗识别能力，由被动识别向主动封堵转化。

2. 加强数据融合能力，强化多元、多模型数据融合能力，打造数据融合网关，强限制数据流转，有条件地开放安全大数据分析能力，更安全、完善地支持相关业务应用；

3. 加强对省分的支持，一方面开放安全大数据分析平台部分能力给省分公司提供欺诈治理支持，另一方面加强信安部对全国欺诈治理的掌控力度，结合新建能力实现全国异常号码的一键智能封堵。

4. 后续研制基于 CNN 卷积神经网络、GCN 图卷积神经网络的基因图谱识别模型，可更有效把控定向诈骗特征，批量抓捕诈骗群体。研究院电信反欺诈研究团队将继续致力于电信诈骗治理新技术的研制工作，发挥创新能力，及时应对层出不穷的通信诈骗新技术新手段。与自主研发，基于电信反欺诈模型，打造成熟人工智能通用能力产品。

3.2.2 恶意网页识别

基于人工智能的诈骗网站识别方案实践

【场景描述】

诈骗网站严重威胁客户的信息和财产安全。360 安全中心发布《2019 年网络诈骗趋势研究报告》显示，2014 年至 2019 年，网络诈骗人均损失成逐年增长趋势，至 2019 年，创下近六年新高。当前诈骗网站已经不局限于通过钓鱼网站盗取用户的隐私数据，色情直播、招嫖、虚假信贷、网络售彩等新型诈骗网站层出不穷，用户深受其害。报告显示，2019 年典型网络诈骗类型包括金融诈骗、游戏诈骗、兼职诈骗、网购诈骗、网络赌博诈骗等，诈骗网站成为实施上述诈骗的重要一环。同时，随着移动互联网的迅速发展和广泛应用，诈骗网站的内容迅速变异、数量持续攀升、行为愈发隐蔽、发展态势日趋产业化，对运营商的检测和处置能力提出了更高的要求。

针对以上严峻形势，中国移动积极践行企业社会责任，利用自身创新和研发能力，基于“主动排雷”的思路，探索了一种基于人工智能的诈骗网站识别应用方案，在高效、准确发现现网中活跃的各类诈骗网站的同时，能够主动拓展发现更多活跃程度不高、隐蔽性更强的诈骗网站，该方案已实现全网的上线应用，持续为客户信息和财产安全保驾护航。

【技术方案】

一、技术方案概述

中国移动提供的基于人工智能的诈骗网站识别方案的整体思路是：从现网数据中获取活跃域名，利用人工智能技术从海量活跃域名中发现和拓展潜在诈骗域名，通过对潜在诈骗域名进行内容爬取及分析，筛选出高度疑似的诈骗网站送交人工审核，对于确认的诈骗域名进行封堵操作。整体系统架构图如图 1 所示，整体系统可以分为数据源、域名发现与拓展、内容取证分析、人工审核、域名封堵五个部分。

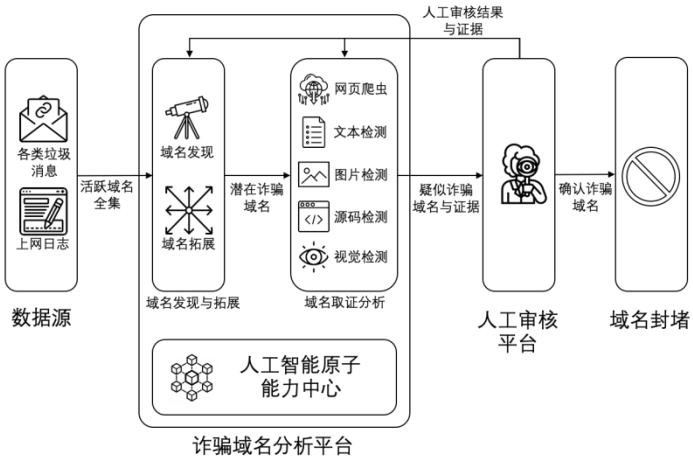

图 1 诈骗网站识别方案架构图

二、域名发现与拓展

1、域名发现

图 2 域名发现模块构成

域名发现模块对网络中存在访问行为的域名进行搜集，并将这些域名信息作为全网的活跃域名全集。活跃域名全集的数据来源包括但不限于用户的上网日志，垃圾短彩信中的域名信息，用户投诉举报数据中的域名信息等。进而通过消息智能分类、域名智能分类、注册备案分析、流量日志分析，从现网活跃域名全集中得到潜在诈骗域名。

2、域名拓展

域名发现模块在活跃域名全集中发现潜在诈骗网站，这意味着域名的发现已经滞后于用户的访问，属于事后发现。域名拓展模块的引入可以弥补域名发现模块的不足，甚至可以在现网监测到用户访问行为前发现潜在诈骗网站，实现诈骗网站的事前治理。

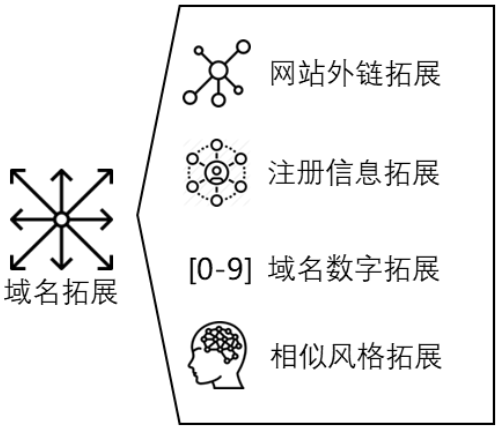

图 3 域名拓展模块构成

网站域名拓展模块能够基于已知的诈骗域名信息进行拓展推广，发现更多与已知诈骗域名存在关联的潜在诈骗域名，由网站外链拓展、备案信息拓展、域名数字拓展、相似风格拓展四个子模块构成。

网站外链拓展：一些赌博类、色情类诈骗网站通常会以友情链接或广告的方式相互链接，方便访问者在网站中跳转。网站外链拓展模块利用这一特征对已知诈骗网站中的外链信息进行迭代爬取，着重爬取诈骗网站中存在的友情链接、广告链接，从而发现更多潜在诈骗域名。

注册信息拓展：诈骗网站制作者往往批量注册大量域名，可以利用域名注册信息的属性关联性来实现域名拓展。注册信息拓展模块通过已知诈骗域名的备案联系人信息反查该联系人注册的其它域名信息，将这些域名作为潜在诈骗域名。

域名数字拓展：赌博类诈骗网站域名中，往往含有大量数字。域名数字拓展模块对存在数字的已知诈骗域名尝试改变数字取值来拓展域名，后续通过爬虫爬取验证可访问后得到潜在诈骗域名。

相似风格拓展：诈骗网站创建者选择域名时会倾向于特定域名风格。相似风格拓展模块使用人工智能技术学习已知诈骗域名的组成风格，利用人工智能的创作能力生成更多与已知诈骗域名构成风格相似的潜在诈骗域名。由于通过人工智能技术生成的域名可能并不真实存在，需要结合爬虫技术来排除掉无法访问的潜在诈骗域名。

三、内容取证分析

内容取证分析模块从潜在诈骗域名的网站内容进行综合分析，从而筛选出高度疑似的诈骗网站，并将相关爬取内容作为判断依据提交人工审核。在爬取网站的过程中，需要爬取网站的文本信息、源码信息、图片信息，另外需要对网站的最终渲染效果进行截图。在得到网站内容信息后，内容取证与分析模块分别从文本、源码、图片、视觉等多个维度对网站进行诈骗分析，并将高度疑似的诈骗网站提交到人工审核平台。

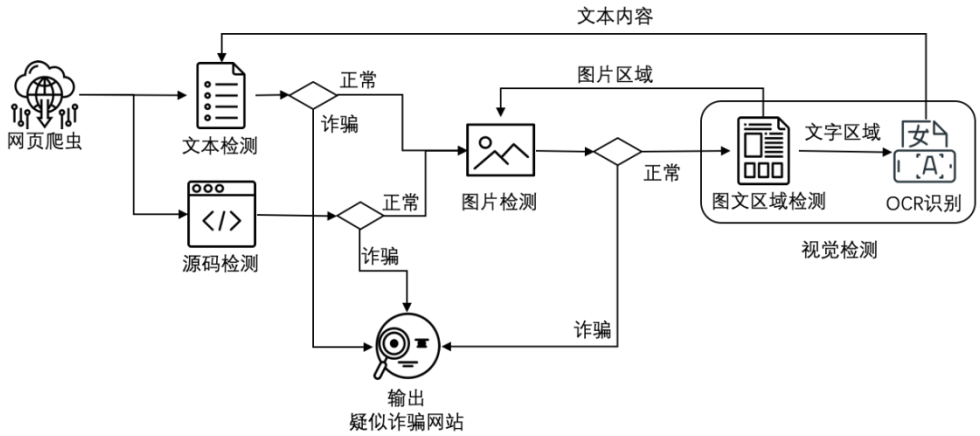

图 4 内容取证分析流程

四、人工审核与特征反哺

人工审核平台审核来自取证分析模块输出的疑似诈骗网站，并通过人工查看证据信息最终确定网站是否为诈骗网站。对于诈骗网站，人工审核平台向域名封堵模块下发封堵指令。经过人工审核判定为诈骗或正常的域名信息会反哺给域名发现与拓展模块和内容分析取证模块中的相关人工智能模块进行更新和矫正。

五、域名封堵

为了能够有效防止用户访问诈骗网址，域名封堵模块将判定为诈骗域名加入拦截名单，当有用户请求拦截名单中的域名时，则中断用户请求，并向用户请求重定向到安全提示页面。

六、分析能力解耦与灵活编排

在整个诈骗网站识别方案中，涉及到大量人工智能分析能力模块。所有能力模块都具有相对独立的功能，且有其单独的进化周期。同时，为迅速适应违规网站的变异衍生，实现对新接入数据源的快速拓展覆盖，诈骗域名分析平台对能力进行原子级解耦，通过对原子化分析能力的组合，实现上层模型的灵活定制，并支持加载外部算法，持续提升监测分析能力，以实现对诈骗网站的新类型、新情况的快速响应能力。

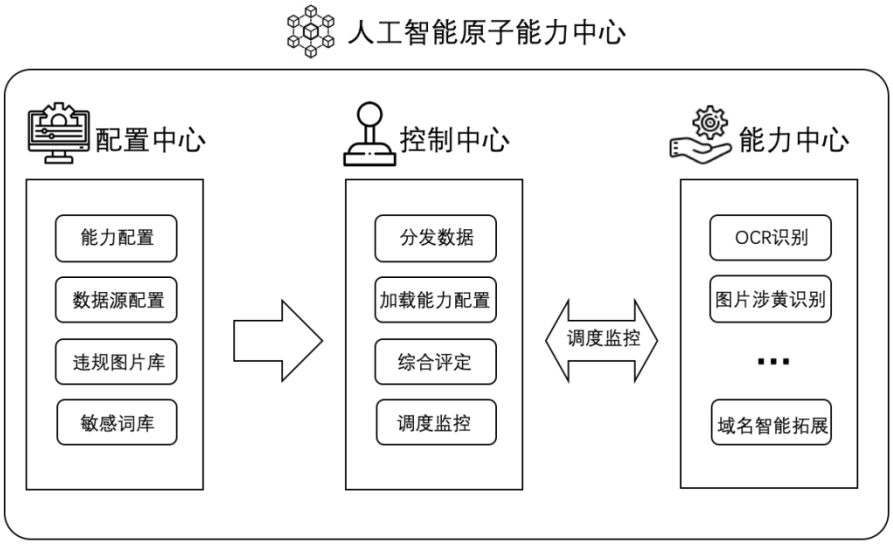

图 5 人工智能原子能力中心架构

【应用效果】

本方案自 2019 年 7 月上线应用，日均分析上网日志、垃圾短彩信、用户举报等超过 200 亿条，月均识别诈骗网站 4.1 万个、阻断 2004.9 亿次诈骗网站访问，有效地保护了客户的信息和财产安全。

【下一步工作建议】

随着生成对抗技术的发展，对抗人工智能技术容易被诈骗者利用，使人工智能分析失效。如诈骗分子可能会使用对抗性的人工智能技术生成诈骗网站域名，防止域名发现模块将域名判别为潜在诈骗域名。另外，随着深度伪造技术成熟，诈骗分子可以在网站中嵌入伪造的图片、音频、视频来骗取受害者信任。

在诈骗网站识别方案下一步工作方向中，将聚焦于如下两个方面：

1） 将着重考虑如何抵御来自人工智能技术的对抗攻击；

2） 将着重如何有效的对深度伪造内容进行识别。

3.2.3 手机恶意软件检测

金踪移动互联网 APP 病毒检测引擎

【场景描述】

随着移动互联网、5G 的高速发展，移动智能设备已经成为人们生活、工作的必需品，各行各业相关相继涌现了丰富的手机软件，这些手机软件在为我们生活的各个方面提供诸多便利，但也为手机恶意软件滋生了发展的土壤。2020 年以来，“新冠肺炎”病毒在全球的蔓延，对全球人民的健康，以及国家经济造成

了不可估量的影响。移动互联网黑灰产、APT 组织，以手机恶意软件为载体，利用新冠肺炎疫情开展网络攻击、以及黑客以新冠肺炎名义进行网络诈骗、勒索、隐私窃取等，进而从中牟取利益。央视 315 晚会还曝光了 50 多款安卓手机软件可能在用户不知情的情况下，偷偷阅读并上传用户短信内容，包括网络交易验证码等行为。无论是关乎到个人隐私信息安全、经济财产安全、还是关系到国家网络空间安全的建设，都需要我们更深层次的认识到手机恶意软件检测的重要性。

传统的恶意 APP 检测手段主要依赖于专家经验，从已有恶意样本中提取检测规则。但是先验的专家规则的效果严重依赖于分析人员的技术水平，对新增恶意样本的检测能力也不尽理想。为了解决传统技术手段遇到的问题，提高手机恶意软件检测的自动化、智能化水平，恒安嘉新基于多年的移动互联网安全防治经验，将 APP 静态分析引擎，动态沙箱技术，与最新的人工智能技术相结合，形成了金踪移动互联网 APP 病毒检测引擎，广泛应用于运营商防治手机恶意软件的现网系统，这里面的技术能力也同时为金融、公安和军队等其他行业赋能。

【技术方案】

金踪移动互联网 APP 病毒检测引擎主要依赖 APP 静态分析引擎提取静态特征，和动态沙箱提取动态特征，通过使用 XGBOOST、深度神经网络等算法对已标注样本的动静态特征进行训练，最终得到基于人工智能的手机恶意软件检测引擎。

静态特征由 55 个行为特征、4 类常见家族特征、以及 15 类病毒常见字符串特征、和 384 个权限特征组成，共 458 个维度。动态特征共 77 个，反映了 APP 运行中的隐私行为、网络行为、文件操作、短信操作等行为特征。静态特征和动态特征共 535 个维度，对每一款 APP 进行了全方位的画像，结合多年沉淀下来海量的黑白样本，和最新的人工智能算法，形成了对手机恶意软件进行自动化和智能化检测的引擎。具体流程如下所示：

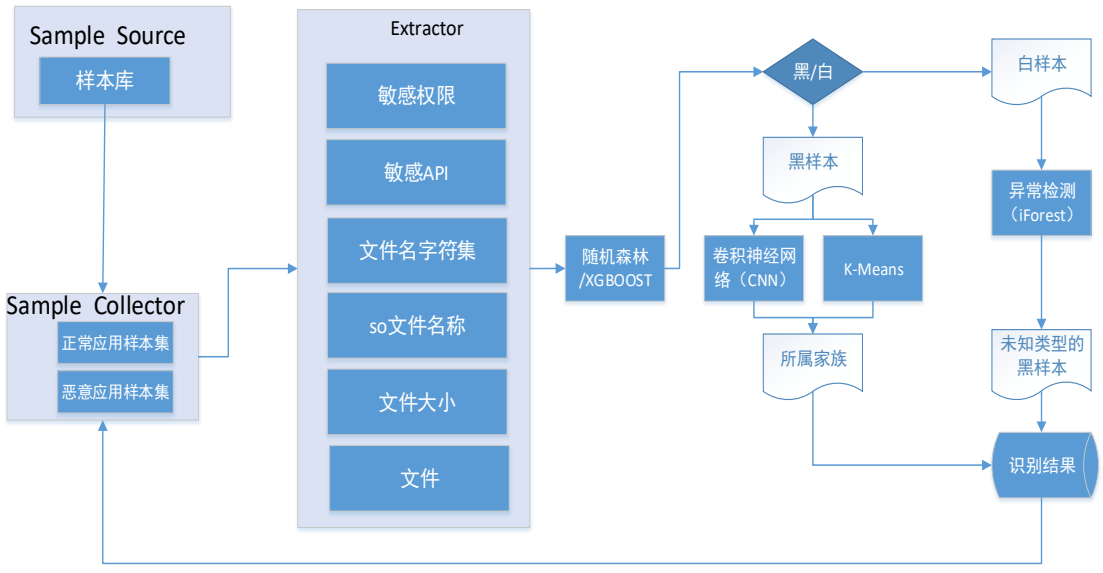

图 1 基于人工智能的手机恶意软件检测流程图

随着整个移动互联网技术的飞速发展，受网络黑色产业链利益驱使，和国家监管层面对手机恶意软件的防治，手机恶意软件往往生命周期较短，更新较快，特征变化较大，离线训练的模型存在失效较快的问题。为了解决这一问题，金踪移动互联网 APP 病毒检测引擎引入模型自更新功能，基于本引擎和第三方检测引擎反馈的最新检测结果对模型的各项指标进行评估，当模型偏效果差达到一定阈值，则会触发模型的自动更新功能，使用最新的标注样本在线重新训练模型并对引擎的模型进行升级。

【应用效果】

以 2020 年上半年为例，恒安嘉新智能创新安全研究院暗影实验室利用大数据舆情监测、移动互联网恶意程序平台、APP 全景态势与案件情报溯源挖掘系统等技术手段，对金踪移动互联网 APP 病毒检测引擎的应用效果进行了详细的数据分析，形成了手机恶意软件态势，具体成果如下：

- 新增移动互联网恶意程序 2118 个（按恶意程序名称去重）
- 新捕获发现样本 1,485,119 个（按照样本 MD5 值去重）

总储备量整体处于稳步增长的趋势。

依据《移动互联网恶意程序描述格式》的八类分类标准，按照移动互联网恶意程序按照分类统计：

- 流氓行为 1, 145, 382（占总数的 77.12%）
- 诱骗欺诈 167, 708 个（占总数的 11.29%）
- 资费消耗 97, 647 个（占总数的 6.57%）
- 系统破坏 42, 400 个（占总数的 2.85%）
- 恶意扣费 16, 161 个（占总数的 1.09%）
- 信息窃取 14, 282 个（占总数的 0.96%）
- 远程控制 1, 340 个（占总数的 0.09%）
- 恶意传播 255 个（占总数的 0.02%）

【下一步工作建议】

监测范围：通过合作运营及监管审查等方式，持续扩大对于通过手机应用商店公开发布的手机软件的监测能力。并利用多种数据来源增加对于非公开渠道发布手机软件的监测能力。

特征向量：对静态、动态沙箱进行优化，提取更多有效特征来区分手机恶意软件和正常软件，增加对手机软件对抗沙箱环境的应对机制，增加对 3D 渲染等新技术的支持，并提高沙箱运行的效率。

算法优化：尝试使用 LightGBM 和深度学习领域最新的人工智能算法，提高单一模型的各项指标。尝试使用多种算法建模，实现模型融合，进一步提高检测效果。

更新标准：《移动互联网恶意程序描述格式》自 2012 年发布以来，移动互联网时代下手机应用程序经历了飞速发展，各类手机软件安全问题不断涌现，需要整个安全行业持续更新手机恶意软件相关标准。

3.2.4 视频行为安全

基于人工智能与机器视觉的视频行为分析系统

【场景描述】

奇安信将人工智能应用在视频行为分析领域，以北京顺诚云计算数据中心为例，该数据中心隶属于北京东方国信科技股份有限公司，总占地面积 2 万平方米，建设机柜规模为 3 千多个。在机房的物理安防方面大量采用了视频监控手段，但基于传统手段的视频监管体系，存在管不全、管不准、管不快的现象，具体问题是：

- 数据中心基本覆盖了对“物”的监控，但对“人”的管控，基本靠自觉、事后查，急需智能化升级。
- 运维人员日常巡检是否准时，交接班、值班等关键人员信息，缺乏记录手段
- 巡检路线是否覆盖关键区域、无死角，无法记录
- 访客进入机房需要工作人员全程陪同，费时费力
- 无法记录访客的行动轨迹，进入非授权区域的危险性很大

【技术方案概述】

基于人工智能与机器视觉的视频行为分析系统将人工智能、机器视觉相结合，落实在物理环境内与人相关的行为安全应用。其主要组件包括高清摄像机、人工智能视频分析系统和决策系统，覆盖了视频行为的感知、理解、分析、决策过程。

本案例所建设系统的功能架构如下：

图 1 基于人工智能与机器视觉的视频行为分析系统功能架构

系统的实施内容包括了在机房出入口、主要通道部署高清摄像机；打通内部 HR 系统，实现人脸数据实时同步；定制化设计平面图，支持轨迹跟踪回放；打通原有视频监控系统，实现视频可检索查询。

方案亮点：

- 预设巡检路线、巡检任务，回放实际巡检轨迹

根据巡检要求灵活设定，针对不同机房、楼层、通道进行设置。规划巡检人员、线路、地点、时间，制定工作计划，设置巡检地点，安排巡检任务。实时查看工作人员所处的地理位置，检查人员到岗情况。使客户对运维人员的管理手段更加智能、可控。

- 重监控区域人脸实时抓拍

在重点区域布控预警，通过人脸属性特征（性别、年龄层、眼镜、口罩、发型）多维分析识别人员身份，一旦发现未登记人员，触发报警，有效帮助客户管理外来人员。

- 目标人员身份识别

根据人员的各种属性特征，将人群归类，如男/女性群体、各年龄段群体。利用人脸+人体特征信息生成个人档案，比单纯基于人脸的识别结果更加准确，且具备自适应、自学习、自纠错能力，自动更新人的档案数据。

图 2 系统在单位出入口进行身份、体温、防疫情况识别的界面

- 行为研判

通过机器视觉技术识别图像中目标的行为，如人奔跑、人离开。

对指定人员（包括未知身份和已知身份人员）时间段内的出现时长\频次进行分析统计。

基于历史数据，利用大数据技术对行为建模，生成历史行为数据，通过对比目标人物当前行为与历史行为找出行为偏差，识别出反常行为。

- 轨迹分析

根据目标人物某段时间内在视频中出现的时间、空间信息，在地图或者区域平面图上绘制行动轨迹。根据目标人物最近时间，在最后一个摄像头内出现的位置，在地图或平面图上定位目标。

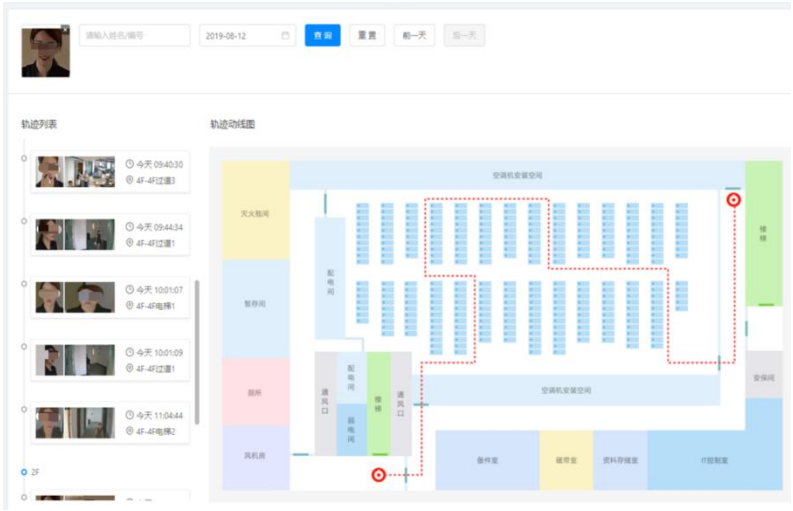

图 3 系统对同一人员行为轨迹跟踪的界面

- 抓拍检索

支持根据图片、目标（人脸、人体）特征属性、时间段、地点等多维度联合检索。

- 全景感知

运用 3D 可视化技术，展示全局安防、人员行为态势，描绘数据流向和变化趋势。

- 人工智能技术的应用

人工智能的应用主要体现在数据处理和人脑智能识别方面，尤其是疫情期间对戴口罩情况下的实人识别技术。

二、核心技术

核心技术一：口罩识别

口罩算法能够判断是否戴口罩以及人脸角度检测，采用了两种方案：

（1）未知人脸位置，直接检测戴口罩人脸，算法模型采用 SDD 类型网络。

图 4 基于 SDD 类型网络检测戴口罩人脸的测试效果

（2）已经人脸位置，判断是否戴口罩，算法模型采用 MobileNet 轻量轻分类网络。

图 5 基于 MobileNet 轻量轻分类网络判断是否戴口罩的测试效果

其中，人脸检测算法推理过程如下：

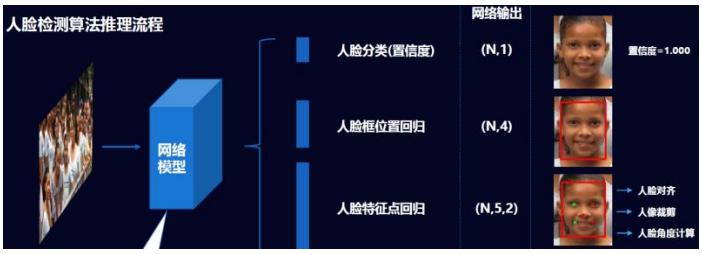

图 6 人脸推测算法推理过程

网络模型设计：

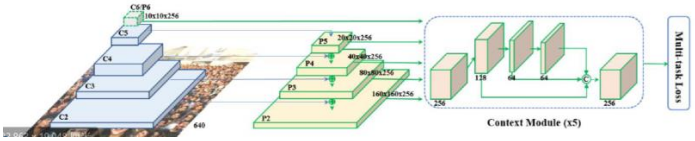

图 7 人脸推测算法的模型设计

核心技术二：人脸识别

人脸识别包括了 1V1 身份鉴别（Identification）、1VN 身份验证（Verification）以及人脸库建立过程。

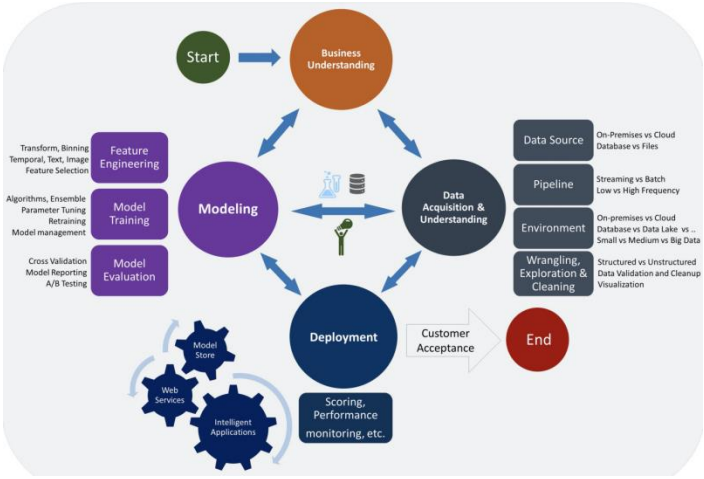

图 7 人脸识别模型训练加载过程

获取存储的结构化/半结构化数据，通过特征工程提取出特征信息，对特征数据进行方差归一化，通过线性回归/SVM/BP 等算法进行模型训练，完成训练后，模型加载到测试环境进行数据集验证，验证完成后更新到线上环境。

【应用效果】

在北京顺诚云计算数据中心和奇安信办公场所等的应用中，基于人工智能与机器视觉的视频行为分析系统能够有效进行重要出入口、重要场所的人员行为分析。至少可以满足疫情防控、安全边界与关口防护、

企业工作效率管控、社会安全治理 4 个方面的需求。人工智能技术大幅提升了疫情时代的人员精确识别难题，机器视觉信息的联动分析实现了大范围、长轨迹下的特定人员跟踪审计功能。

【下一步工作建议】

该方案可用于疫情防控、安全边界与关口防护、企业工作效率管控、社会安全治理等领域，可向企业管理、部队管理、学生公寓、医院等人员数量较多、对人员行为、轨迹有安防要求的场所推广。

3.3 数据安全篇

3.3.1 数据分级分类

数据分级分类

【场景描述】

在大数据应用日益广泛的今天，数据资源共享和开放已经成为促进大数据产业发展的关键，但由于数据的敏感性，加之各行业数据分类分级标准的滞后和缺失，使数据开放和共享面临诸多困难。通过 AI 算法进行自动化数据分级分类，有利于稳步推进数据开放和共享，为大数据发展应用奠定基础，实现数据价值的最大挖掘利用。

数据实施分级管理，能够进一步明确数据保护对象，有助于企业组织合理分配数据保护资源，是建立健全数据生命周期保护框架的基础，也是有的放矢实施数据安全管理的前提条件。

同时，统一的数据分级管理制度，能够促进数据在机构间、行业间的安全流动，有利于数据价值的充分释放。

图 1 基于人工智能机器学习的敏感数据识别分级分类

Use Case1：数据发现，梳理数据资产

数据资产梳理是数据安全的基础。知道企业究竟有多少数据、这些数据在哪里、有哪些类型的数据、有哪些是敏感数据，这些数据的敏感等级分别是什么？只有明确了保护的目标，才能针对安全风险进行有针对性的防护。

政务数据共享交换这项业务中，各类单位与组织会把数据资源集中到大数据局的数据资源平台。那么对于大数据局来说，首先要做的一项工作就是进行数据发现，通过对资产的全面盘点，形成相应的数据资产地图，知道自己手里有什么之后，才能有针对性的保护数据资产安全。

AiGuard 基于网络嗅探技术，可自动寻找发现网络环境中存在的数据库；在定义 IP 地址段后，通过端口扫描的方式，发现各类数据库，表和字段。

Use Case 2：数据发现，安全性检查

数据库漏洞、弱口令、错误的部署或配置不当都会很容易让数据库陷入危难之中。全面检查和测试数据库是非常有必要的，以确保数据安全。

AiGuard 能够充分扫描出数据库系统的安全漏洞和威胁并提供智能的修复建议，通过对数据库进行全自动化的扫描，从而帮助用户保持数据库在最佳安全健康的状态。

Use Case3：敏感数据分布，自动识别敏感数据分级分类

敏感数据分布是资产梳理的关键一步，只有明确敏感数据资产都有哪些、在被哪些部门、哪些人员如何使用，才能真正保证数据在使用中的安全。

AiGuard 内置模板规则和 AI 算法，通过自动识别敏感数据、分级分类以及权限梳理，让用户轻松掌握数据库及其数量、表数量、敏感表数量、敏感字段数量、敏感数据类别和敏感数据使用情况等信息。

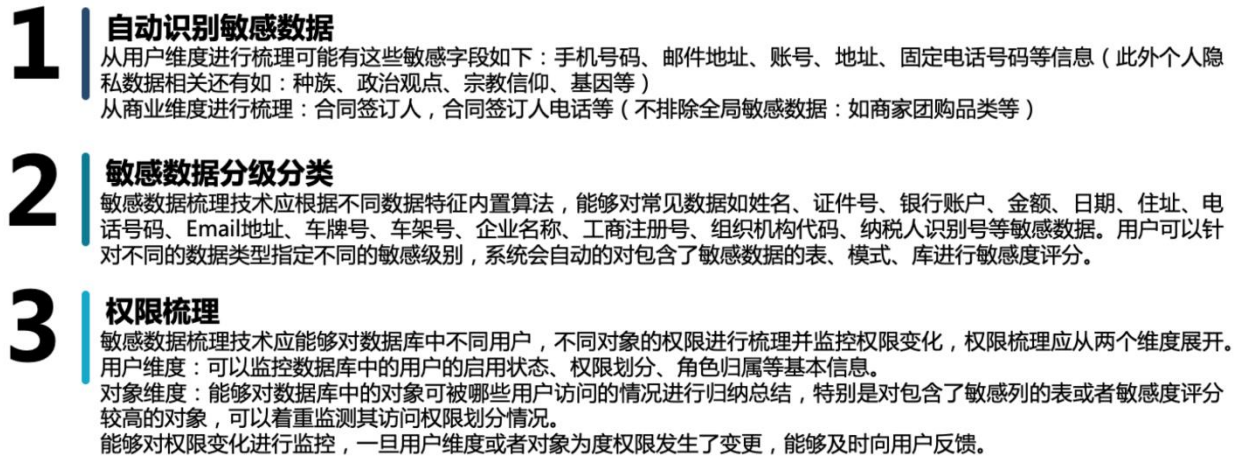

图 2 AI 自动识别生成敏感数据分布

Use Case4：权限最小化/细粒度授权管控

数据面临的访问权限问题主要有：员工被赋予过多的超出其工作所需的权限；反之，则是没有开启足够的权限；另外，权限还可能被恶意使用。

例如在医疗行业，由于好奇产生的越权操作问题在医院是广泛存在的客观现实。譬如医生会因为对某种疾病感兴趣，进而查询和阅读非授权病历，DBA 会因为碍于情面帮助朋友查询他人隐私数据。

再例如，企业数据仓库中有一张关于用户的基本信息表 User，其中有手机号 mobile、姓名 username 两个字段。我们在划分数据安全层级的时，将用户 mobile 的安全等级划分为 L2 要高于 username 的等级 L1，并规定只有访问权限达到 L2 的运营部门才能访问 mobile 字段。这样在公司各个部门需要访问注册用户基本信息表 User 时，我们只需检查访问者是否来自运营部门，如果是运营部可以访问 mobile，如果不是只能访问 username 信息了。这样就有效的防止用户手机号被不相关工作人员泄露出去，同时也不影响查询用户 username 的需求。

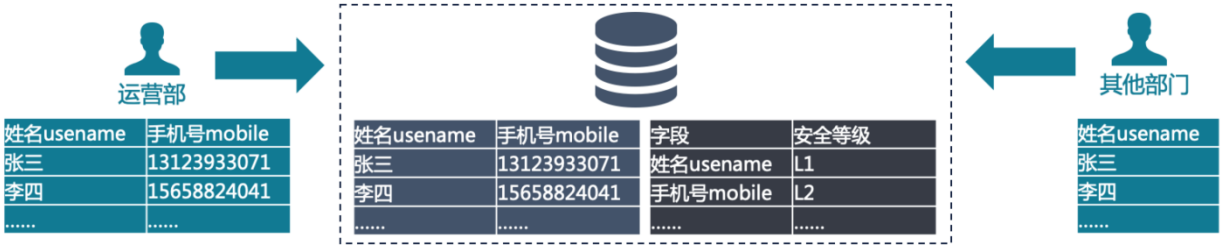

图 3 根据敏感数据级别细粒度权限控制

Use Case5：数据脱敏

在系统开发测试过程中，由于要高度模拟生产环境，因此很多情况下，需要使用生产环境中的生产数据进行系统开发测试。而生产数据一旦流转到开发测试环境，其数据的安全性则无法得到有效保障。由此，需要通过数据分级分类识别出需做安全保护的敏感数据，再用脱敏技术确保数据中的敏感信息被漂白，但又不影响开发测试人员对于数据的使用。

通过建立数据脱敏机制，对发放到开发测试环境的生产数据预先进行脱敏处理，确保经过脱敏后的数据不再带有敏感信息，且数据面向开发测试人员可用。

图 4 数据脱敏

Use Case6：避免个人信息泄露

社会上出现了大量兜售房主信息、股民信息、商务人士信息、车主信息、电信用户信息、患者信息的现象，并形成了一个新兴的产业。比如，个人在办理购房、购车、住院等手续之后，相关信息被有关机构或其工作人员卖给房屋中介、保险公司、母婴用品企业、广告公司等。

例如火车票、网购订单中根据数据分级分类情况加以不同策略的脱敏处理

【身份证号】显示最后四位，其他隐藏。共计 18 位或者 15 位，比如：**********1234

【中文姓名】只显示第一个汉字，其他隐藏为 2 个星号，比如：李**

【地址】只显示到地区，不显示详细地址，比如：上海徐汇区漕河泾开发区***

图 5 个人信息脱敏保护

Use Case7：数据分类存储和备份

将数据分类分级，以便你知道一些关于它的基本事实，例如文件内有什么，敏感数据有哪些，为什么创建它，谁创建的，谁应该能看到它谁不应该等等，这一批不小的信息决定了数据该如何被处理和存放。如果它是公司的重要信息，你可能需要多次备份，加密并设置访问权限。如果它是公司团建活动的计划，可能就不需要太严格的措施。

对不同类别、级别的数据采取对应的物理或逻辑隔离措施。同时，在存储过程中要采用最稳定的架构，建立授权和最小权限机制，建立实时备份机制，建立多重索引机制，建立数据使用追溯机制和多地冗余备份机制。 建立起与数据量级规模相当、范围适度、多地保存的数据备份机制，结合系统运行状态制定容灾备份策略和规程，恢复范围和目标、切换规程、灾后恢复运行操作指引。同时，定期组织进行灾难恢复的教育与培训，确灾难性情况下数据可提取，可恢复。

Use Case8：数据定期消亡

实现数据定期消亡。为了保证数据和个人隐私安全，在数据失效后，依照相关法律法规要求应建立相应的数据销毁机制，明确销毁方式和销毁要求。同时，遵守全过程可审计原则，建立数据销毁策略和管理制度，明确销毁数据范围和流程，记录数据删除的操作时间、操作人、操作方式、数据内容等相关信息。

例如：《信息安全技术 个人信息安全规范》标准中规定了开展收集、存储、使用、共享、转让、公开披露、删除等个人信息处理活动应遵循的原则和安全要求。 其中明确指出，在符合特定情形时，应及时删除个人信息，实现日常业务功能所涉及的系统中去除个人信息（敏感数据），使其保持不可被检索、 访问的状态。

Use Case9：数据开放共享

政府数据的分级由数据的敏感程度划分。政府数据的分级结果是数据开放和共享的依据。分级结果将确定该类型政府数据是否适合开放和共享、数据开放和共享的范围，以及在对该级别政府数据进行开放和共享前是否需要脱密和脱敏（包括逻辑数据运算等处理方式）处理等。

根据《政务信息资源共享管理暂行办法》的规定：

【共享属性】

政务数据资源的共享类型包括：无条件共享、有条件共享、不予共享三类。值域范围对应 1、2、3。

【开放属性】

数据资源面向社会开放的属性，包括“是”和“否”，对应取值分别为 1 和 0。

	政府数据敏感程度		
	非敏感数据	涉及用户隐私数据	涉及国家秘密数据
等级划分	公开数据	内部数据	涉密数据

数据等级	数据等级管控要求
公开数据	政府部门无条件共享；可以完全开放。
内部数据	原则上政府部门无条件共享，部分涉及公民、法人和其他组织权益的敏感数据可政府部门有条件共享；按国家法律法规决定是否开放，原则上不违反国家法律法规的条件下，予以开放或脱敏开放。
涉密数据	按国家法律法规处理，决定是否共享，可根据要求选择政府部门条件共享或不予共享；原则上不允许开放，对于部分需要开放的数据，需要进行脱密处理，且控制数据分析类型。

图 6 政府数据敏感程度和数据等级管控要求

【技术方案】

强大的数据识别能力

内置基于深度学习+条件随机场算法的（BiLSTM+CRF）实体识别模型，可准确识别人名、地名、机构名称、时间、日期、金额等各类实体内容；内置基于正则匹配+luhn 校验的复合识别算法模式，用以精确识别身份证、银行卡等敏感信息；支持自定义创建正则、字典、算法三类识别规则。

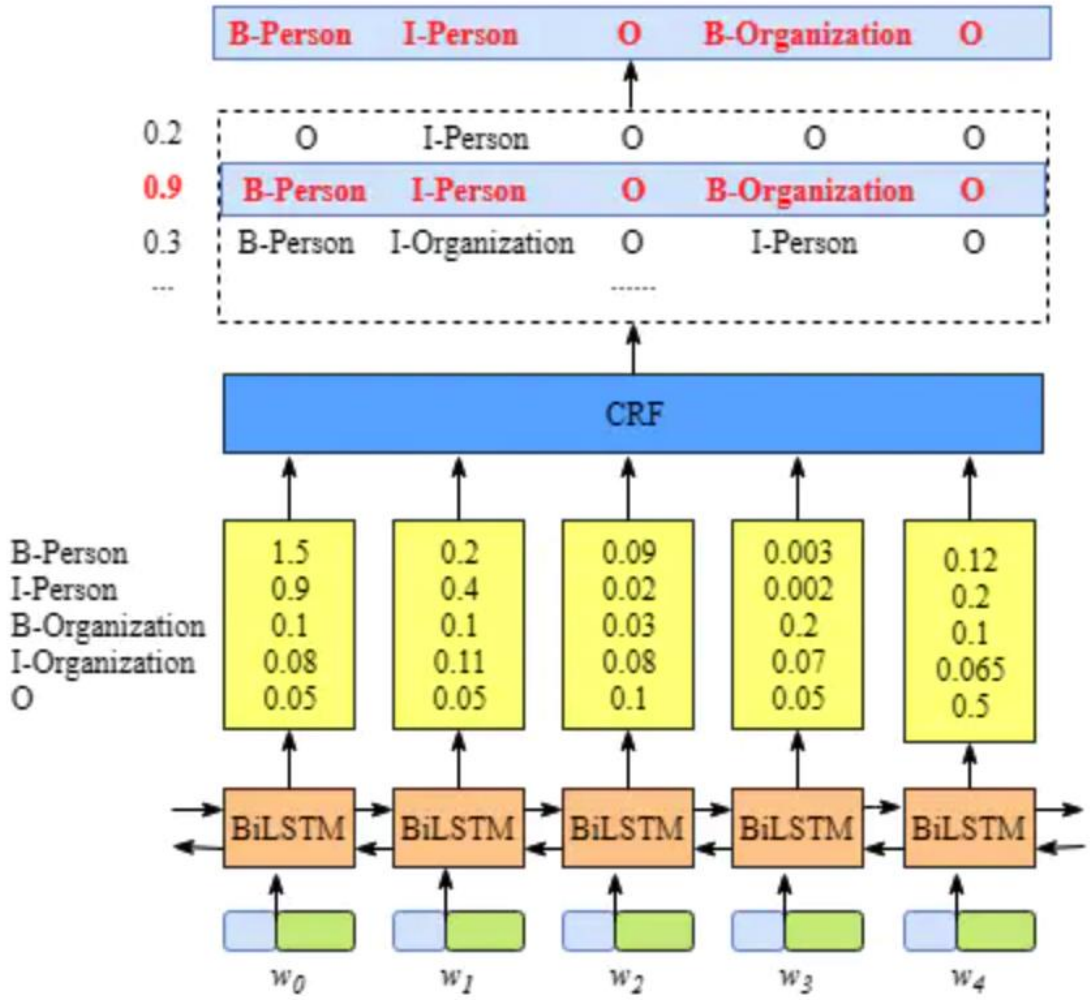

图 7 实体识别模型示意图

丰富的分级分类包

数据分类可以根据不同维度进行划分，如：业务分类、行业分级分类指南等；鉴于此种情况，分级分类系统内置多种分类标准的分类分级包，满足不同行业不同客户的需求；分级分类包如下所示：根据识别规则（规则 ID），识别到特定字段，结合分级分类包详情，给出推荐的分级分类列表。

类别	类别编码	字段名称	数据等级	规则ID
交易-投资者管理-投资者基本信息-个人投资者基本信息	A-4-1-1	投资者姓名	3	1
交易-投资者管理-投资者基本信息-机构投资者基本信息	A-4-1-2	法人代表	2	1
公司运营管理-综合管理-人力数据-一般人员信息(公开)	D-2-1-1	员工姓名	1	1
公司运营管理-综合管理-人力数据-档案管理	D-2-1-3	员工姓名	3	1

图 8 分级分类包

精准的分类分级推荐

如图 8，根据规则与分级分类包可能给出数据类别有多个；此时，需要使用 Apriori 或 FP-growth 关联规则算法，训练出推荐模型，给出分类分级推荐度排序，最高推荐度作为缺省分类、分级；用户也可以根据推荐度列表手动选择数据类别、等级。

【应用效果】

安恒分级分类系统，目前已在金融、政务行业有过良好的实践，内置包括金融、政务、医疗、通信等行业的分级分类包，以及部分法规涉及数据的自定义分类（如：个人隐私数据保护规范），其中数据识别准确率达 84%，分类推荐准确率达 71%。

【下一步工作建议】

下一步，分级分类系统将在以下两个方面有所提高：

1、数据识别准确率 95%以上，分类准确率突破 80%

2、支持更多的数据识别技术，包括 OCR 图片识别、音视频识别等；能够对用户全量数据资产进行分级分类、为满足网络安全法、数据安全法等法规提出数据保护要求打下坚实的基础。

敏感数据智能识别及分级分类

【场景描述】

过去 20 年间，经济形态发生了很大的变化，呈现为数据资产形态的无形资产逐渐超越了有型资产成为了资产的主体。2015 年标准普尔 500 企业中无形资产占到总资产的 84%。而 15 年前仅为 32%，而 IT 系统中存储流转的数据则是企业的核心资产。这些数据有的可以被贩卖，有的直接关系到用户的资金安全，有的还会损害企业甚至国家的信誉和利益。因此，对于数据的重视和保护成为 IT 领域新的关注焦点。

Ultra-DSM 是神州泰岳安全公司发布的数据安全管理系统，面向企业的数据安全管控、数据安全监控、数据安全运营的平台。

【技术方案】

一、背景描述

- 海量的数据背景下，人工手动进行分类效率低下。在企业数据分类工作中，通过人工分类往往需要几个月甚至更长的时间才能完成，而且在分类过程中，又会有数据新增或者旧的数据发生变更，造成数据分类工作无法准确、及时的交付
- 人员业务知识有限，无法专业的对不同业务数据进行归类。数据分类人员一般需要具有专业的业务知识，分类过程需要企业调配相应的业务资源进行持续的配合，需要大量的时间和沟通成本。

而通过文件名、文件格式进行分类，很难保证分类与内容的准确匹配。

- 数据分类后，对于不同种类的数据需要形成对应的安全检测策略。随着数据量和数据内容的不断变更，还需要企业花大量的人力物力进行安全策略的更新，同样给企业带来更多的资源消耗。

二、数据内容分析

使用国际一流的数据内容分析引擎，对结构化和非结构化数据的内容进行准确识别，并依据不同数据内容的不同特征，将数据按照不同分类粒度进行归类，辅助人工专业介入，快速对海量数据进行快速分类。同时，根据分类结果，结合企业内部分级标准，对数据进行分级整理。

三、智能学习组件，构建数据分类分级策略模型

基于复合型指纹技术，对结构化数据和非结构化数据进行指纹扫描，根据指纹的结果生成指纹库。在安全检查时做到相似度匹配。自动化持续指纹任务可以做到对目标数据的持续分析，目标数据发生变化，相应的指纹库也会对应处理，保证检测的准确性。

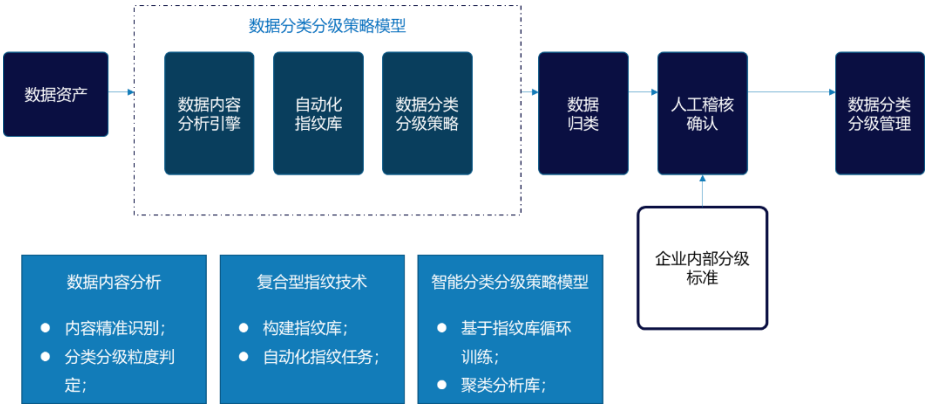

图 1 数据分类分级策略模型

使用智能学习组件，对不同类别、级别的数据分别进行机器学习，生成学习结果共安全策略使用。同时，智能学习和可以按照要求进行定时、定量的持续循环工作。保证安全策略的检测内容随时保持最新状态，和企业的数据资产内容保持匹配。

【应用效果】

数据安全落地的过程中，数据分类分级和安全策略一直是安全落地的关注点，也是自愿投入资源最多的部分之一。和传统安全的关注点“威胁”不同，数据本身是在不断变化的，而变化存在不确定性，而且是常态。作为安全管理者，在资源有限的情况下必须实现针对数据的持续、智能的有效监控。

1. 文件智能分类

数字化转型带来的最大的变化是从数据资产的角度去看，很多我们过去被忽略的、没有收集的数据都变成了数据资产。不依赖于数据的表面属性，根据数据的真实内容进行分类。能够有效的帮助数据所有者和安全管理者从数据的业务属性进行分析，并通过不同业务数据的共性进行归类。能够将传统人工数据分

类的时间缩短 90%以上。而且，随着大数据时代的到来，数据的分类分级往往要从源头抓起。结合下面提到的策略更新和数据标签功能，能有效的将数据分类分级落到实处，并且实现动态可持续。

2. 智能数据智能学习

在真实的企业数据环境中，无论是结构化数据，还是非结构化数据，都在持续的更新和变化中。在企业形成了初期的数据安全基线后，保证数据安全基线能够随着数据资产的变化不断调整，也是考验数据安全工作有效性的重要指标。智能学习功能，能够对分类后的数据进行定期、自动、智能的持续分析。数据安全基线初始化以后，提供就可以回特定的分类后数据进行持续的内容分析，包括增量分析、定期分析等，并在此基础上不断的更新智能分析的结果。接下来安全策略就可以调用这些结果，对数据的存储、使用和传输中的数据进行实时监测，保证安全策略的持续有效性。

3. 数据指纹智能识别

对于一些特别重要的核心数据资产，如红头文件、战略计划、网络信息、核心业务数据等，其内容的全部和部分都必须纳入监控范围。智能指纹学习功能，能够将核心数据资产已数据指纹的方式进行存储。在不影响保密等级和范围的情况下，完整的进行保护。通过建立指纹任务，对重要资产进行智能、持续的学习，形成数据指纹库。策略通过调用指纹库，对送检数据进行检测，保证重要数据资产内容安全。

【下一步工作建议】

持续丰富整体智能分类分级策略，并构建多维分类分级维度，提供交互式灵活数据运营、管控功能，输出更真实数据管控能力。

同时针对 5G 时代，面向更多更新数据类型、数据内容，不断丰富、扩展数据分类分级策略模型，适应更新的发展。

3.3.2 数据风险评估

基于用户行为的数据安全异常检测

【场景描述】

目前多数企业已经规划或开展信息安全管控策略及措施的实施，但在数据安全保护层面的措施仅限于传统网络安全、存储冗余、备份及集中化管理、桌面安全管理等层面，对于数据安全领域关注度不够，在核心数据资产的使用、传输、保管、销毁的过程中存在较多安全风险，同时也加大了信息安全管理工作的难度，所面临的关键问题及风险统计如下：（1）数据资产不清，梳理难度大；（2）数据共享缺乏统一管理，泄露风险大；（3）数据合规性风险；（4）数据安全管理风险。

针对上述关键问题，绿盟提出基于用户行为的数据安全异常检测技术，把注意力放在特定用户的活动上，通过多种统计及机器学习算法建立用户行为模式，当黑客的行为与合法用户出现不同时进行判定并预警，从而发现数据泄露风险。

【技术方案】

较为完整的系统结构如下：

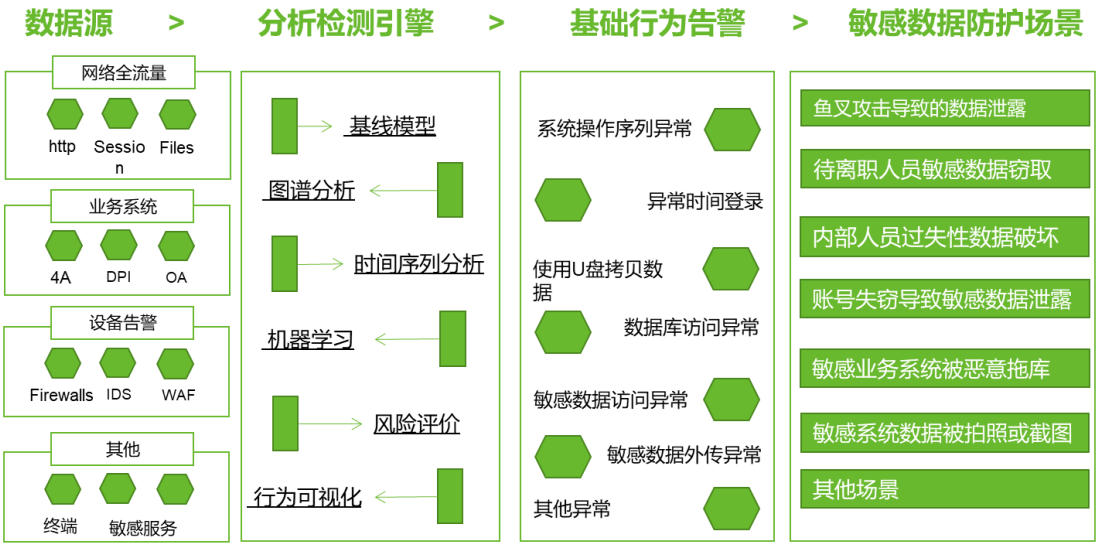

图 1 用户数据访问异常识别架构

- 数据源：主要来源于多源异构数据，包括网络流量监测数据，对会话、协议、文件的解析日志；业务系统数据，登录系统日志、文件访问日志等；相关全设备告警；数据防护设备日志等；
- 分析检测引擎利用多源数据采集和安全分析，构筑以用户/网络实体行为为视角的分析能力。
- 基线行为告警按照 5W1H 分析模式，利用自学习的行为基线算法，训练生成动态的数据行为基线模型；
- 敏感数据防护目前支持多种场景的数据异常的检测，基于正常行为模式进行异常行为检测及比对，有效发现内网用户横向攻击和违规操作、以业务用户者身份入侵到内网，利用内部人员身份盗取/变更数据等。

基于机器学习技术，采用以用户/实体为中心的分析方法，运用数据模型和规则，对用户和实体的行为描绘，形成模型如下：

- 基线模型：以人、资产、数据之间的历史行为操作数据为输入，按照 5W1H 分析模式，利用自学习的行为基线算法，训练生成动态的数据行为基线模型；利用动态行为基线做检测，当数据行为发生改变或者偏离基线时发出告警。

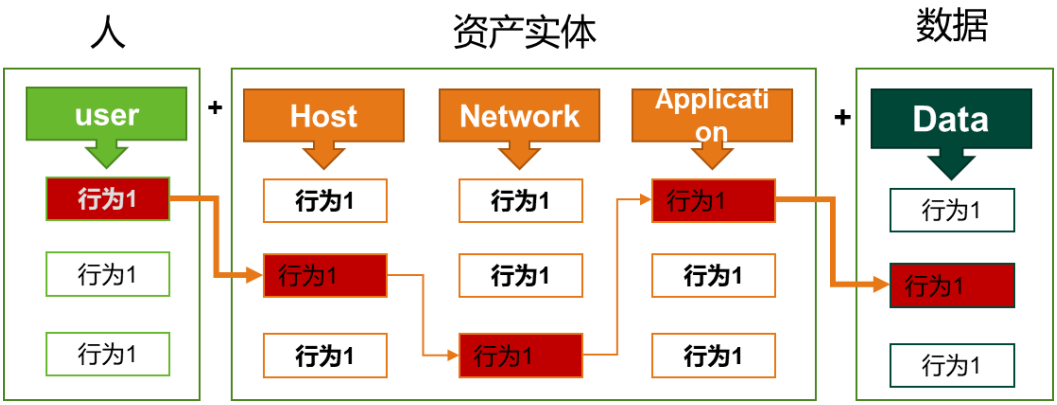

图 2 行为分析模型

- 时间序列模型：基于时间序列分解的异常行为分析发现和提取行为中序列突发成分;然后基于向量之间的欧几里得距离,用遍历和匹配方法提取周期子序列;最后将行为序列分解为突发成分、周期成分和随机成分，重点分析突发行为及周期行为，已确定其是否为异常行为，如周期性的非核心时段超量下载文档，周期性超量上传核心文档到互联网，我们都可以认定为内部人员窃取数据。

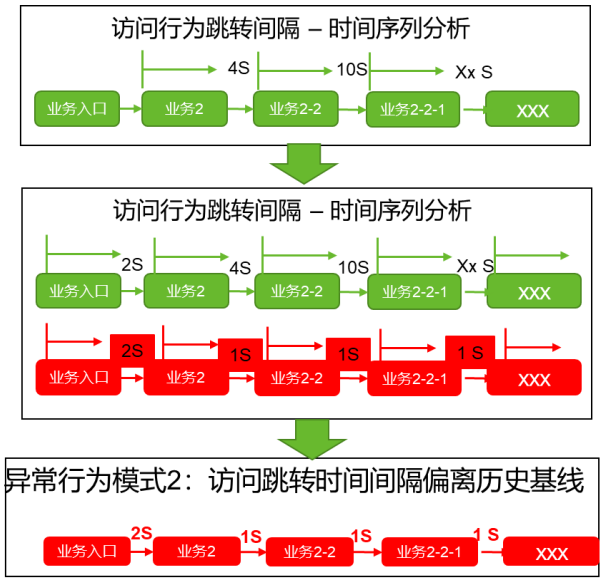

图 3 数据异常访问模型

- 图谱分析模型：基于历史的可信访问行为提取访问规则，利用图连通性算法进行行为聚类，形成可划分的访问行为簇，从访问者的角度提取出可访问基线，从被访者角度提取被访问基线，并可视化呈现。让管理者对于敏感数据访问情况，由一无所知转变为可视可管。

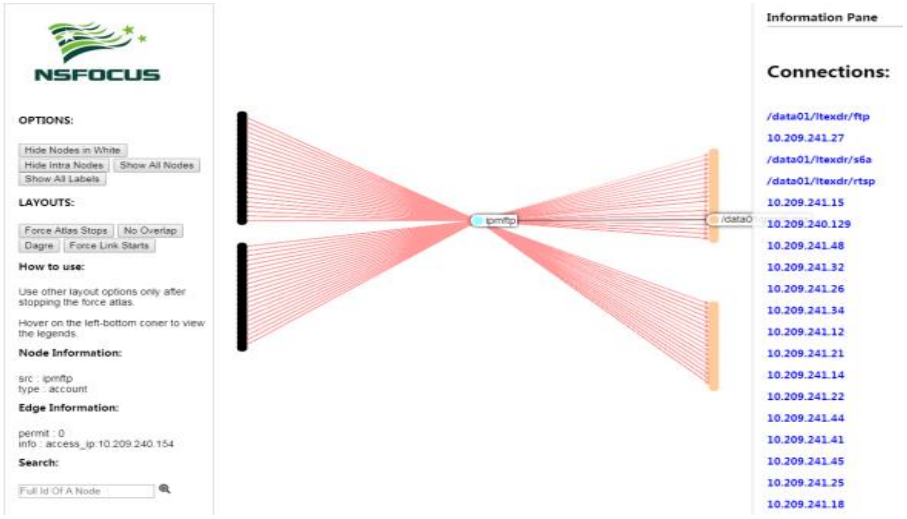

图 4 图谱分析

【应用效果】

经过多次实践分析，目前支持多种场景的数据异常的检测：包括数据泄露检测、账号失陷检测、绕过控制行为检测和非法外联检测等。

【下一步工作建议】

对数据的可视化监控、风险点排除，及时预警、及时阻止对数据的非法使用行为，覆盖更多的检测场景。

3.3.3 数据防泄漏

基于语义的敏感文档识别

【场景描述】

随着互联网的迅猛发展，网络对于社会的影响越来越大，成为了信息传播的重要渠道。与此同时，信息过载、网络内容安全、信息泄露等问题日益突出，安全保密工作面临的形势更加严峻。为尽可能减少敏感数据的外泄对企事业单位、甚至国家安全和利益造成的威胁，必须对敏感数据进行严格管控。因此，如何有效、快速识别敏感数据就成了需要解决的重要问题，而关于敏感数据识别技术的研究对于防止敏感数据外泄与增强泄密隐患的发现能力具有重要的意义。敏感数据的表现形式多种多样，包括文本、图片、视频、音频等。本案例关注的是最常见的一种敏感数据，也就是文本形式的敏感数据检测，即敏感文档识别问题。

传统的敏感文档识别技术主要基于关键词表与词频统计，将文档中是否出现关键词及出现的数量作为敏感文档识别的主要依据。然而，现实中有很多场景中不适用这一方法。具体来说，在一些具体应用场景中，会预先指定一些文档为敏感文档，需要检测识别与这些指定文档语义相近的所有文档。这些指定的敏感文档并不一定是一般意义上的敏感文档，可能不包含特定的敏感词，而只是包含一些公司内部敏感信息，比如内部会议纪要等。针对这种需求，核心问题是计算不同文档之间的相似度，通过与指定敏感文档相似度的高低来判断任意一份文档是否敏感。常用的方法是对文档中出现的所有词进行统计，然后比对前 N 个高频词表作为判断依据。但是这一方法不能准确捕获文档的语义信息，因为同样的语义可以通过不同的用词与语句表达。

对于以上问题，启明星辰发挥创新、研发优势，设计研发了一套基于文本建模的数据防泄露系统。该系统使用语义层次分析技术，能够识别出与指定敏感文档语义上相近的文档，从而识别出敏感文档，继而避免敏感数据的泄露和传播。

【技术方案】

本技术方案的主要流程如下图所示。首先，对待检测文档与敏感文档进行文本预处理，从中抽取文本内容并进行分词，再应用词嵌入模型将其转换为词向量序列。词嵌入模型可直接选用一些已有的中文预训练模型（例如腾讯在 2018 年公开的中文预训练模型，包含 800 万个词或短语，词嵌入向量维度为 200），也可以通过使用具体领域的相关语料进行预训练得到。然后将待检测文档的词向量序列与敏感文档的词向量序列应用词移距离算法计算得到两者的相似性度量，从而得出最终识别结果。

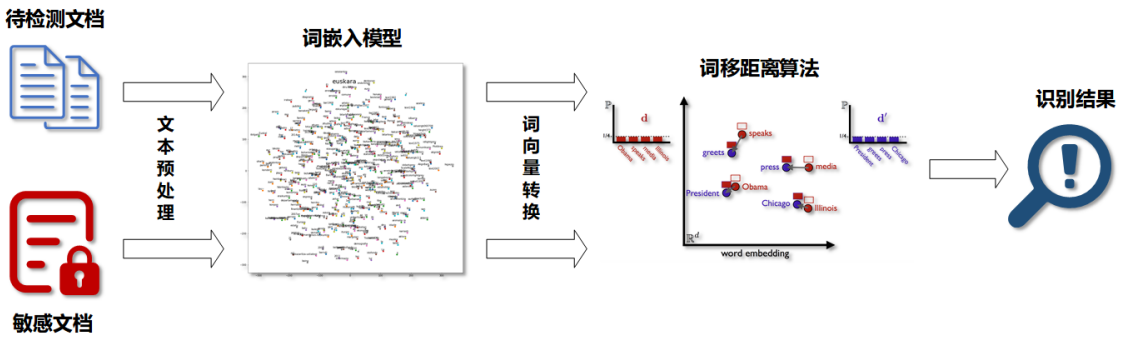

图 1 敏感文档识别技术方案示意图

这一技术方案的核心是在文档相似度计算中采用了基于语料库算法中的无监督学习方法——词移距离算法。所有词都可以通过 Word2vec、GloVe 等方法映射到词嵌入向量空间中；对于任意两段文字，其中一段中的每一个词都能移动到另一段中的某一个词，这样就可以计算出每一对词在词嵌入向量空间中移动的距离，所有词对的距离之和就是这两段文字的词移距离。在此基础上，为两个词的对应关系定义了基于词嵌入向量的权重，两段文字的加权词移距离称为文本距离。算法主要通过最优化方法计算任意两段文字的最优文本距离，作为文档相似性的度量依据。

【应用效果】

相比于前述的传统方法，采用本方案的敏感文档识别技术能够在保证准确率不降低的前提下检测发现更多的语义相近的敏感文档。具体来说，在一次实际测试过程中，总计处理约 5000 份文档，相比于传统方法，能发现的敏感文档要多 4%（约 200 份）。

【下一步工作建议】

优化算法实现，进一步提升计算效率，以适用于数据量大的场景。另外，本技术方案所采用的是无监督学习的方法，其总体准确率仍有改善提升的空间。在下一步工作中，将设计有监督机器学习的方法，以进一步提升敏感文档的识别准确率。

3.4 业务安全篇

3.4.1 物联网

人工智能赋能物联网安全防护

【场景描述】

物联网系统是一种虚拟网络与现实世界实时交互的新型系统，给人们带来方便的同时，也带来了信息暴露的危险。随着物联网的迅猛发展，安全问题不可避免地摆在面前，从对各种可能的攻击进行分类可分为：对感知层的攻击、对网络层的攻击、对应用层的攻击。

感知识别层的设备、节点等无人看管，容易受到物理操纵。物联网多用来代替人完成一些复杂、危险和繁琐的工作。在此种情况下，物联网中设备、节点的工作环境大都是无人监控的。因此，攻击者很容易就能接触到这些设备，从而对设备或其嵌入其中的传感器节点进行破坏。攻击者甚至可以通过更换设备中的软硬件，对它们进行非法或者破坏性操控。感知层主要面临的威胁有：软件漏洞、数据泄露、恶意软件感染、非法入侵、非安全通信、服务中断。

物联网要求对感知数据能够进行安全的传输，系统节点多、数据量大，传输要求高。同时物联网的传输往往需要在异构的网络之间进行，物联网在信息传输中多使用方便快速的无线传输方式，暴露在外的无线信号很容易成为攻击者窃取和干扰的对象。大量信息外流会对物联网的信息安全产生严重影响。攻击者也可以在物联网无线信号覆盖的区域内，通过发射无线电信号来进行干扰，从而使无线通信网络不能正常工作，甚至瘫痪。网络层主要面临的威胁有：DDoS 攻击、信令风暴、非法入侵、数据泄露等。

物联网使得人们随时随地都可以方便快捷地获取物品的确切位置及周围环境等相关信息，这造成潜在的隐私威胁。例如， RFID 标签能被嵌入任何物品中，一旦被嵌入而物品的使用者未察觉，物品的使用者将会不受控制地被扫描、定位及追踪。应用层主要面临的威胁有：DDoS 攻击、入侵攻击、恶意代码、数据泄露、业务欺诈、系统漏洞。

针对上述问题，恒安嘉新研发基于人工智能的物联网安全防护系统，包括数据采集层、数据处理层、业务支撑、数据服务层和展示层等。

【技术方案】

物联网安全检测服务平台采用大数据处理架构，将多源异构数据采集技术，与机器学习、大数据、新人工智能技术等技术相结合，构建沙箱研判、隐蔽信道检测、攻击行为建模分析等能力，以海量异构数据深度实时分析为基础，形成物联网安全事件的全面检测重点平台，提供物联网设备的安全检测和安全风险预警能力。

一、关键技术

物联网安全检测与态势感知平台：集成了物联网卡人工智能安全模型技术， 基于机器学习的实时流量识别技术，基于人工智能技术的高级持续威胁预警技术、**基于海量接入的安全威胁溯源技术**。

物联网卡人工智能安全模型技术：挪用异常检测模型、滥用异常检测模型、异常流量/短信模型。

基于机器学习的实时流量识别技术： 实现“5°空间画像”，即：攻击轨迹、网络资产、流量占比、文件、失陷资产。系统通过 5°空间画像实现了安全态势可视化、流量回溯、深度分析、威胁分析、样本分析、资产管理等功能。

基于人工智能技术的高级持续威胁预警技术：系统采用了静态特征、动态行为、威胁情报和机器学习等多种检测手段，从 Cyber Kill Chain 的多阶段进行“断链式”检测 APT 攻击。系统的沙箱引擎从用户交互差异性、运行环境差异性、业务逻辑差异性三方面实现了 200 种以上的反逃逸技术，如替换操作系统所有和虚拟机有关的指纹、模拟网络、模拟用户对系统的使用痕迹等。

基于海量接入的安全威胁溯源技术：物联网万物互联海量接入的设备，引入海量的数据和安全问题。通过构建安全模型， 深度挖掘分析能力，借助人工智能和机器自学习能力，勾画网络攻击的全景图， 并追踪溯源，精准定位安全威胁的源头。

基于人工智能的物联网安全防护系统功能架构图如下所示：

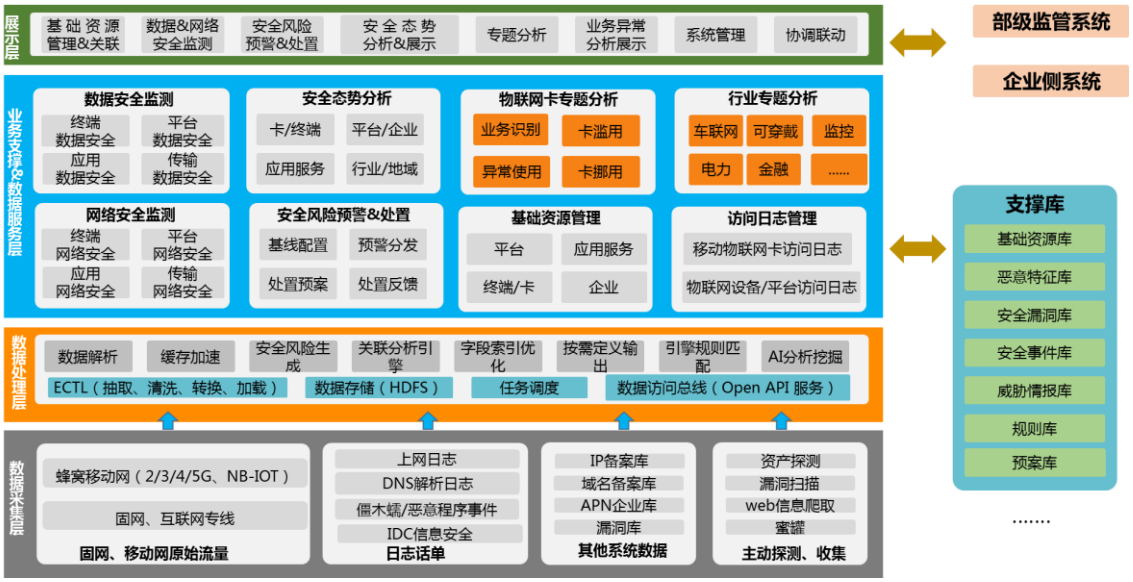

图 1 物联网安全防护系统功能架构图

物联网安全防护系统主要由业务采集层、数据处理层、业务支撑&数据服务层、展示层四大部分建设组成。按照“集中存储安全数据，持续扩充分析场景”的方式建设一个覆盖物联网全业务和海量终端的平台。物联网安全防护系统提供基础资源管理、指令管理、标识管理、查询管理、安全监测、APP 监测、安全预警、安全态势分析、知识库管理、资产画像、业务监测、资产发现与核验、专题分析、报告管理、系统管理等功能。主要特点：物联网异接入构海量数据，通过人工智能算法学习大量筛选数据训练，构建完美安全检测和感知系统，及时发现系统的已知威胁和潜在威胁， 提供准确及时的安全威胁信息和应对策略。

物联网安全防护系统各层如下。

1) 多异构数据采集层：海量接入系统内部数据，包括物联网卡，固网、移动网络原始流量，日志话单，主动探测收集数据， 漏洞库， 域名解析库等等。

2) 可扩展的物联网安全数据中心：采用多样的、可适配数据源的方式对物联网安全相关数据进行采集、清洗、标准化、存储，提供离线、实时、全文检索等多种数据订阅及分析方式，数据能力具备线性扩展能力。

3) 开放式物联网安全分析架构：基于高质量的安全数据，物联网安全防护系统提供数据开放，各应用可向物联网安全防护系统订阅数据用于自身的分析需要，逐步扩充其他行业卡和物联网产品的安全分析模型和应用，最终实现全业务覆盖。

4) 可视化物联网安全态势：实现物联网综合安全态势，并对物联网安全态势、威胁预警和风险通告以图形化展示，可为各类用户分配账号，提供直观、强大、清晰的安全风控和预警能力，以及重大问题、事件的整体性报告。

【应用效果】

2020 年 3 月，系统在湖南进行了试点部署，通过分析 15 天的数据，识别物联网卡 6.1 万张，捕获了 4.3 万次攻击事件，识别物联网协议 28 种。

2020 年 4 月，系统在广东某运营商进行了部署。当日识别异常访问政府网站的公用类物联网卡 5 张，滥用的导航类物联网卡 3 张，异常抄表类物联网卡数量 104 张，DOS 攻击行为物联网卡数量 82 张，可疑诈骗卡 53 张。此外系统监测到异常基站 10 个，存在安全风险的信息传输事件 307 条。

2020 年 6 月，系统在江西省试点部署，10 日分析事件约 28 亿条，发现活跃物联网卡数量 130 万张，识别省内物联网平台 1606 个，相关机构及企业 178 个。

【下一步工作建议】

结合试点和实验情况，后续系统需要在如下几方面提升能力：

1) 分析研判的效果与监测的覆盖率密切相关，需逐步推进监测分析的覆盖范围；

2) 异常访问的特征和识别方式还需要不断升级，虽然当前识别算法已可就常见的主要物联网异常进行识别，但仍有与行业和特色物联网服务的异常未能识别；

整体系统在威胁共享和事件处置方面的能力还需要结合行业规管建设逐步完善。

3.4.2 工业互联网

人工智能赋能工业互联网安全监测服务平台

【场景描述】

制造业数字化转型之势如火如荼，工业互联网对制造业数字化转型的驱动能力正逐步显现，这直接催生了工业互联网安全监测与安全服务诉求。工业互联网安全与传统的工业控制系统安全进一步融合，传统的工业控制系统在引入了工业以太网和 TCP/IP 等开放性通信协议，系统平台趋于开放化和标准化，与外部网络的连接变得更为紧密与频繁，特别是工业 4.0 概念融合智能工厂、智能生产、智能物流的业务特征，传统安全风险和网络攻击通过互联网侵入到工控系统中，在能源、石油化工、水利、核工业、交通等领域中安全攻击事件屡见不鲜。在整个数字化转型建设中，主管部门、行业监管单位和工业运营企业对属地内工业资产分布、联网设备现状、安全风险状态、安全事件监测与处置面临着难以有效感知、难以及时监测、难以有效处置的直接困惑。

针对上述问题，恒安嘉新借助人工智能技术构建监测和服务一体化手段，进行异构多源数据融合、协助威胁识别和风险研判、辅助安全事件处置，赋能工业互联网安全。

工业互联网安全监测服务平台基于主管部门业务布局，为各级主管部门提供面向属地内工业互联网安全监测服务，辅助科学决策，助力科技监管。

工业互联网安全监测服务平台通过集中部署方式，为区域或集团属地内行业用户提供工业互联网安全监测和威胁处置服务，适用于工业制造、能源、交通运输、医疗、轻工业生成等工业互联网安全监测与防护场景。

工业互联网安全监测服务平台可部署在企业出入口，为工业企业提供暴露资产监测、安全事件监测、数据泄露监测、异常流量监测等服务，并提供本单位整体安全态势，适用于工业企业自建防护手段场景，为工业企业的安全运营保驾护航。

【技术方案】

工业互联网安全监测服务平台基于多源异构数据采集技术汇聚基础数据、活跃数据和威胁信息，通过**智能化数据中台**技术满足各类业务数据接入、标准化存储、智能化分析和自动化运维，以响应各场景用户的业务关切为核心，具备贴近省级工业互联网安全监测场景下的分析研判和溯源处置功能，为用户提供“摸清家底”、“知晓风险”、“溯清源头”、“快速处置”、“科学决策”的体系化支撑能力，利用人工智能技术赋能科学决策，助力科技监管。

一、关键技术

为实现对联网暴露设备监测、工业互联网安全事件等新场景下的安全监测与态势感知，平台集成了深度安全 NTA 分析技术、基于机器学习的流量识别技术、基于机器学习的风险监测技术、基于海量威胁情报威胁溯源等技术。其中：

深度安全 NTA 分析技术：基于主被动相结合的方式，从工业企业外网和内网两个视角实现工业互联网安全监测和防护评价。深度安全 NTA 设备支持精细粒度协议分析，支持 400 多种协议分析能力。支持不同部署环境的流量识别和分析，通过松耦合架构结合灵活部署方式，支撑复杂的应用场景，并通过场景建模等技术手段实现性能优化。采用基于网络、协议、编码、流的组合特征，能够识别多类型工业互联网平台应用，具备规则特征的快速集成，支持识别规则的在线升级。

基于机器学习的流量识别技术：通过机器学习技术，在不依赖于端口和规则特征的流量识别前置条件下，从数据流中提取的特征属性，构建工业互联网安全监测业务分类模型并完成流量的识别。相关特征属性主要包括数据包特征和数据流特征，数据包特征主要体现在数据流内的数据包大小、数据包间隔时间以及数据包速率等；数据流的特征则主要体现在数据流的传输层协议、源端口、目的端口、流持续时间、流传输的数据量以及含有标识位(FIN,SYN,RST,PUSH,ACK,URG) 的数据包个数等。

基于机器学习的风险监测技术：基于传播渠道的行为检测技术，基于工业互联网企业、平台、设备、业务的基础特征数据设计机器学习模型，分析异常数据传输和敏感数据泄露事件。集成暴露资产监测、异常流量监测、安全威胁监测技术手段，依托风险检测引擎的协同工作，可针对恶意攻击、脆弱性利用、Web 漏洞、主动数据泄露行为等场景开展检测和研判，建立多维度的特征扫描机制，构建 SHA1、软件包名、类名字节码、特征字、资源名、软件签名、字符串等多态特征广谱扫描机制和方法，提高对潜在风险的识别能力，实现对安全风险多态特征进行精准快速定位。

基于海量威胁情报的威胁源溯源技术：依托快速处置业务要求，追溯工业互联网恶意软件发布源头，通过构建深度挖掘分析能力、充分发挥互联网信息钻取能力，实现对重点传播源渠道的主动、定向追踪。

二、技术架构

工业互联网安全监测服务平台汇聚各业务系统的数据，具备主动资产探测能力，部署在企业出入口、网络节点等网络环境中，开展属地内工业互联网安全检查和管理业务。

平台架构如下图：

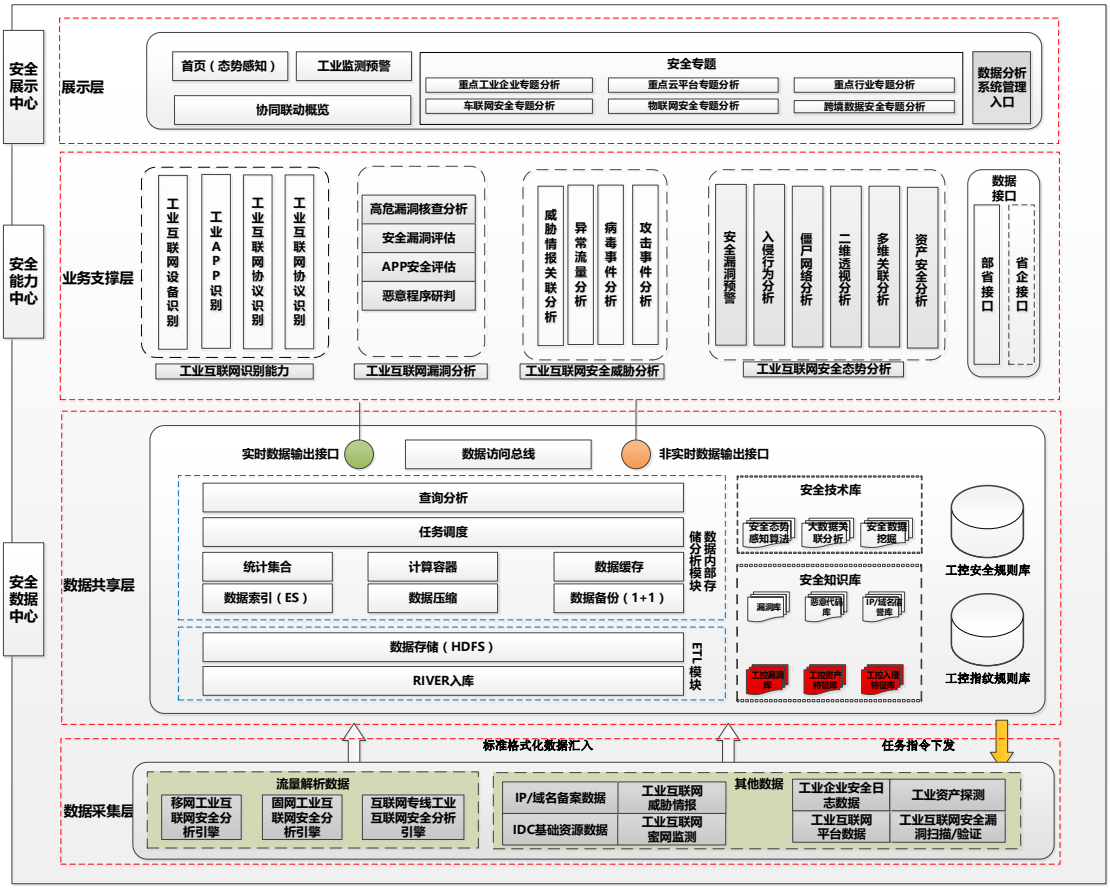

图 1 平台架构图

产品架构从底向上：

数据采集层：数据采集是工业互联网安全监测服务平台的基础，主要通过在移动互联网接入口、城域网出口，互联网专线，工业互联网平台等主要网络位置部署流量采集与深度包检测设备，采集基础数据和业务数据。通过接口方式汇聚流量和日志信息，基于主动探测和蜜罐诱捕手段收集威胁情报/安全事件数据，对数据进行归一化融合处理。

数据共享层：提供统一的数据存储功能，对接入的数据进行存储、分析，形成基础资源库，集成安全态势感知算法、安全数据挖掘算法形成安全技术库，通过与其他子系统的业务交互和处理积累知识库和规则库，为面向网络侧的安全监测和安全分析能力提供业务数据支撑。

业务支撑层：基于统一总线（实时数据输出接口和非实时数据输出接口）接收业务数据管理共享层的处理数据，依托安全分析和监测模块的协同工作，实现工业互联网联网识别（联网/暴露设备识别、工业 App 识别、工业互联网协议识别）和业务支撑（安全监测、威胁预警、协同联动）等功能。

展示层：将工业互联网安全态势及工业资产监测结果统一呈现，包括态势感知、工业监测预警、安全专题等主要前台界面。该层提供业务交互的入口，实现协同联动和后台的业务交互。

三、业务流程

（1）识别对象

首先，通过接口汇入和手工导入的方式汇集主动数据，然后对监测对象进行识别，识别出工业互联网平台、联网设备及系统、工业应用协议和安全事件，用以安全分析的数据输入。

（2）安全分析

其次，将识别分析结果与主动探测数据（包括蜜罐诱捕数据）、威胁情报信息（安全漏洞库）做融合处理，进行深度安全分析，包括资产漏洞分析、入侵行为分析、敏感数据传输分析、威胁情报关联分析在内的多维度分析处理，形成多维研判分析能力。

（3）能力支撑

最后，将研判分析输出的安全业务支撑能力（包括威胁预警、处置协同和信息共享能力）依托于信息共享能力做横向扩展，威胁预警能力支撑网络安全通报业务，处置协同用以支持应急协作业务，汇总两个业务方向的数据，形成整体的安全态势。

【应用效果】

1、资产识别

进行联网资产识别和分析。通过流量监测支持 Siemens S7、Modbus、IEC 60870-5-104、DNP3、EtherNet/IP、BACnet、Tridium、 Niagara Fox、redlion-crimson3 等主流通信协议识别，能够对 HTTP、DNS、FTP、SMTP、HTTP2 等 15 种以上互联网协议识别。被动手段可从流量中识别 150 种小类的工业相关软硬件产品，能够提取设备类型、厂商、版本号等信息；主动手段可探测 100 种以上工业通信协议，支持 300 种以上联网软硬件产品小类的识别。

2、风险监测

平台以漏洞识别和安全事件监测为抓手，具备安全事件、恶意网络资源、恶意程序和安全漏洞发现能力，能够面向属地内工业资产、工控设备、关键设备、重要工业基础设施（应用系统）开展安全监测和预警。可监测工业互联网中僵尸、木马等恶意程序导致的主机受控事件，以及恶意扫描、异常流量、漏洞利用、Web 攻击、拒绝服务攻击、暴力破解等成功入侵安全事件；能够从安全事件中半自动化研判恶意 IP、恶意域名、恶意 URL，形成恶意资源库；支持监测计算机恶意程序、移动恶意程序、工控恶意程序、物联网恶意程序及其他恶意程序，已积累百万级规则；积累数万条漏洞规则，能够发现 Web 漏洞、数据库漏洞、操作系统漏洞、PLC 设备漏洞等常见漏洞类型。以某省单月为例：监测到某省内暴露在公网安防监控

设备 71 台，工业控制设备 17 台，涉及交通运输、食品、教育等相关行业的管理系统 12 个，且多个系统存在弱口令。

3、安全服务

基于平台积累的知识库，并汇集各渠道监测的海量样本数据为基础，通过构建深度挖掘分析能力，充分发挥互联网信息钻取能力，实现对重点传播源渠道的溯源分析。以某省为例，2020 年 1 月至 3 月，监测攻击次数 31479 次，涉及 17 家单位，发现成功事件 25 次，以此面向属地内交通运输、能源、制造业等行业提供监测预警服务。

【下一步工作建议】

面向监测业务拓展需求，平台应继续落实人工智能赋能建设任务，面向业务下沉到工业企业维度，打造快速处置能力，辅助工业企业科学决策，安全运营。

1、人工智能辅助风险处置

平台聚焦工业互联网安全监管需求，以安全事件为抓手，抽取业务场景，通过安全监测引擎和威胁分析引擎的协同，构建快速有效的威胁处置机制，并针对有明显通用特征的安全监测结果进行深度挖掘和多维分析、研判，实现面向通用事件和个体事件精准研判和快速处置能力。后续，平台面向数据安全监测，依托机器学习技术，实现对工业数据分类分级自动化分析和归类，辅助工业企业、工业互联网平台企业进行数据异常传输和数据泄露风险处置。

2、人工智能辅助科学运营

平台面向属地内业务需要，通过平台+运营模式，协助客户从监管侧、企业侧、管道侧出具安全分析报告，辅助用户科学决策。依托安全事件监测，面向属地内工业企业开展安全预警和风险通报服务，协助客户落实属地内工业企业的主体责任。后续，面向企业主体，借助人工智能算法实现对资产、威胁、事件和业务的量化指标，生成企业的整体安全指数，明确企业整体风险级别，协助科学运营。

3.5 终端安全篇

云端联动的诈骗短信智能分析与检测

【场景描述】

近年来，电信网络诈骗违法犯罪行为日益猖獗，不法分子通过群发各类诈骗到用户手机，诱导受害人访问钓鱼诈骗网址、或下载手机木马、或拨打诈骗电话，最终实现对受害人财产的非法占有。

典型的诈骗短信样例：

- 电话诈骗短信：“*尊敬的建行用户！您的信用卡已达 100%提额标准。请致电 021-61521192 进行办理，我行将在一个工作日完成额度调整【中国建设银行】*”

- 木马诈骗短信：“*【违章查吧】您的爱车在本月有违章行为,请进 https://m.clcxwz.cn/进行查询.已处理请忽略退订回 T*”
- 钓鱼诈骗短信：“*【车主通知】您的资料已过期，为避免影响您的使用，请及时登录 ww.bjetci.com 更新信息。退订回复 TD*”

针对此类与真实业务短信非常相似的诈骗短信，目前各类手机安全软件、垃圾短信拦截系统在识别与拦截能力能上存在不足。本方案，基于部署在中国移动自有营业厅的“守望者”终端安全工具箱，通过集成在工具箱中的诈骗短信人工智能识别模型，结合云端联动的诈骗电话库、木马网址库、钓鱼网址库，实现对被检测用户手机终端上诈骗短信的自动识别，改善用户体验，提升运营商企业形象。

【技术方案】

本方案整体系统架构如下图所示：

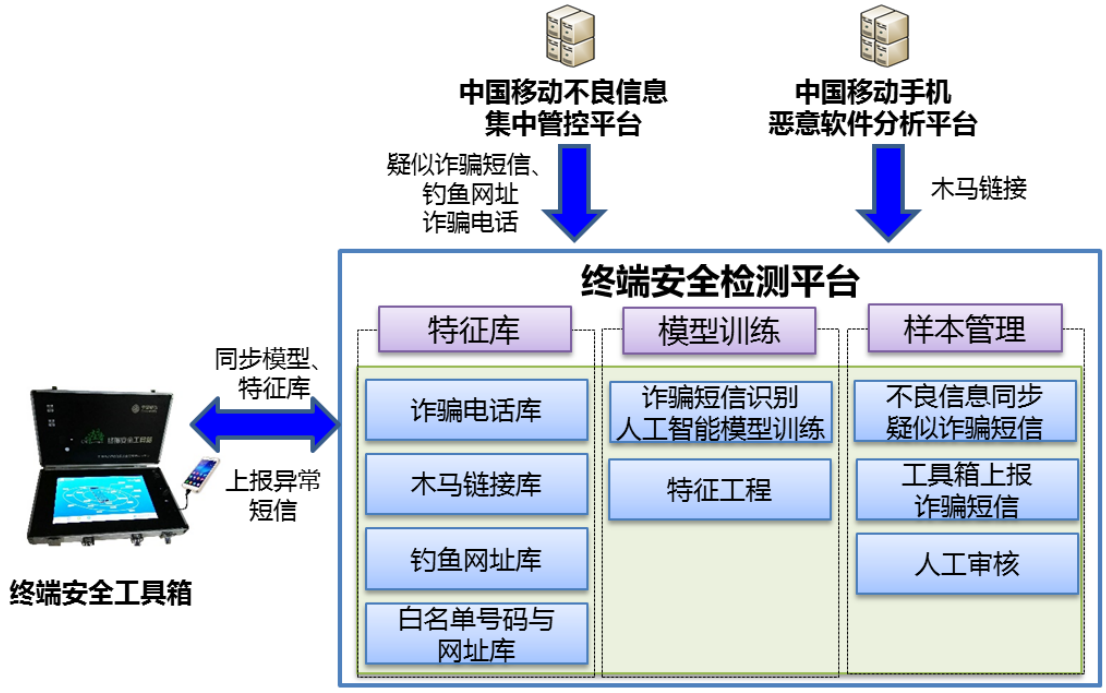

图 1 云端联动的诈骗短信智能分析方案

整体方案包括几个部分：

- 特征库管理：云侧的终端安全评测平台，从不良信息集中管控平台、手机恶意软件分析平台同步诈骗电话库、钓鱼网址库、木马链接库，并通过安全专家分析真实业务短信，梳理白名单号码与网址库，生成对应的特征库数据，同时支持终端侧的终端安全工具箱定期从云端获取最新的特征库数据；
- 模型训练：云端的终端安全评测平台，利用从不良信息集中管控平台同步过来的疑似诈骗短信、工具箱检测到的用户手机终端诈骗短信作为训练样本，基于支持向量机（SVM）算法，结合安全

专家人工提取特征、标记样本分类，训练诈骗短信识别人工智能模型。训练出来的模型，可以部署到终端安全工具箱上，并支持后台定期更新；

- 样本管理：云端的终端安全评测平台，存储从不良信息集中管控平台同步过来的疑似诈骗短信、工具箱检测到的用户手机终端诈骗短信，以满足人工智能模型训练所需。同时，平台支持针对工具箱上报疑似诈骗短信的人工审核确认；
- 诈骗短信识别：终端侧的终端安全工具箱，从云端下载特征库数据、诈骗短信人工智能识别模型，基于特征指纹（诈骗电话、钓鱼网址、木马链接的哈希计算值）、人工智能识别模型对用户手机侧的短信进行识别，向手机用户发出预警，并在经用户同意后将识别到的诈骗短信上报到云端平台。

模型训练：基于 SVM 算法的诈骗短信识别人工智能模型

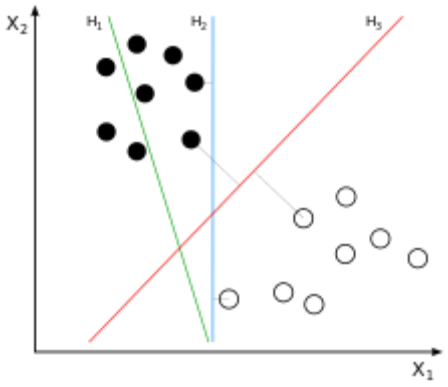

图 2 支持向量机（SVM）算法

支持向量机（Support Vector Machine, SVM）是一种监督学习（supervised learning）算法，被公认为是文本分类中效果较为优秀的一种方法。它是一种建立在统计学习理论基础上的机器学习方法。该算法基于结构风险最小化原理，将数据集合压缩到支持向量集合，学习得到分类决策函数。这种技术解决了以往需要无穷大样本数量的问题，它只需要将一定数量的文本通过计算抽象成向量化的训练文本数据，提高了分类的精确率。

本方案中，首先会将利用从不良信息集中管控平台同步过来的疑似诈骗短信、工具箱检测到的用户手机终端诈骗短信进行诈骗短信分类标记。

然后，平台会将样本短信数据作预处理后，依据特征工程定义的特征，生成特征向量。这里的特征，主要包括安全专家梳理的特征词集（比如：友情、提醒、建设银行、信用卡、提额、办理、请拨、专线、回复、调整，通常会覆盖所有已知类型诈骗短信的关键字）、是否存在非白名单之外的电话号码、是否存在非白名单之外的 URL 地址、短信长度。

示例短信：“尊敬的建行用户！您的信用卡已达 100%提额标准。请致电 021-61521192 进行办理，我行将在一个工作日完成额度调整【中国建设银行】”，在经过特征计算后，可以生成特征向量：

[0，1，0，。。。。。。。。，1，0，53]

最后，平台基于训练样本的特征向量、分类标签，使用“Python 语言+Sciket-Learn 机器学习框架+sklearn.svm.SVC 分类算法”，训练诈骗短信人工智能模型，并生成 PMML（Predictive Model Markup Language, 测模型标记语言)）模型文件，可以部署到终端侧的工具箱（JAVA 语言）上使用。

诈骗短信识别：基于指纹匹配和诈骗短信识别人工智能模型的综合研判

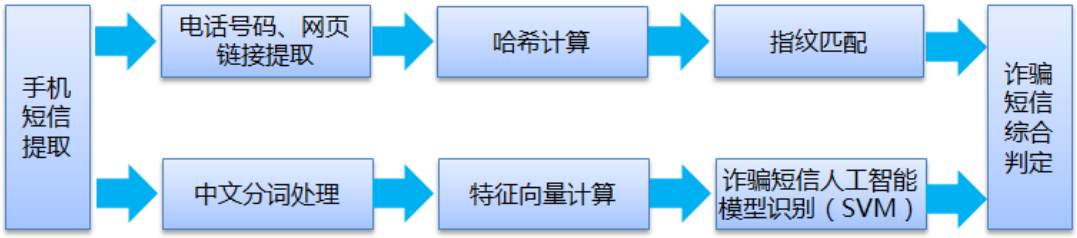

图 3 诈骗短信研判

本方案提出了基于指纹匹配和人工智能模型（SVM）识别的诈骗短信综合研判流程：

1. 首先，终端侧的终端安全工具箱，提取用户手机侧包含 URL 链接和电话号码的各类短信；

2. 然后提取其中的 URL 链接、电话号码，计算哈希值，然后再与白名单、黑名单进行匹配，获取匹配类型

3. 再者，对短信进行分词处理，依据特征词集，生成特征向量，并调用训练好的诈骗短信研判模型，获取识别分类结果

4. 最后，依据指纹匹配结果、文本分类结果，给出是否诈骗短信的综合判定。

【应用效果】

本方案在 2018 年底中国移动终端安全技检测平台四期上线，累计在全国 145 家营业厅展开试点。截止 2019 年底，共计发现各类诈骗短信 15 多万条，涉及用户终端 8 万余台，大大提高客户感知。

【下一步工作建议】

模型持续更新：由于诈骗短信的样式不断翻新，方案需要及时跟踪业务变化，依据不良信息集中管控平台同步过来的疑似诈骗短信、工具箱检测到的用户手机终端诈骗短信的最新变化，持续更新研判模型。

新业务场景拓展：随着 5G 消息的上线，后续需要考虑如何实现对用户手机接收到的诈骗类 5G 消息研判。

4 总结与展望

综上，人工智能融于安全、人工智能赋能安全、人工智能重塑安全，网络空间安全防护势必将成为人工智能应用规模发展、新型智慧社会美好发展的重中之重，世界将最终走向万物智能时代。一方面，随着

云、管、端、边等基础设施的融合协同，特别是在以 5G 为首、人工智能为核的“新基建”的加持下，万物互联和数据汇聚持续加速，大大降低了人工智能技术引入和应用门槛，全面推动人工智能深度融入经济社会发展。另一方面，随着全球网络空间安全协作逐渐达成共识，相关政策法规和安全标准完善统一，以及计算机视觉、机器学习、自然语言处理、音视频识别等人工智能技术的不断进步，引入人工智能技术是提升网络空间安全防护水平的必然要求。

本案例集主要从全球人工智能技术演进与战略法规布局、人工智能赋能安全的内涵与意义等出发，围绕通信网络安全、内容安全、数据安全、业务安全和终端安全等五大应用场景，详尽梳理收录了 30 篇人工智能赋能安全的典型案例和优秀实践。其中，通信网络安全 18 篇、内容安全 5 篇、数据安全 4 篇、业务安全 2 篇、终端安全 1 篇。整体来看，人工智能技术在传统的通信网络安全领域发展已较为成熟且全面，在内容、数据、业务和终端等领域的布局和融合尚需进一步推动和完善。总体来说，人工智能不仅可以解决现网安全难题，还可以进一步深化和发掘人工智能的潜在应用。根据 MarketsandMarkets 公司 2018 年发布的《安全市场中人工智能》报告，全球人工智能赋能安全市场规模在 2017 年已达 39.2 亿美元，预计 2025 年将达到 348.1 亿美元，平均每年增加率超过 30%。可以预见，后续人工智能在安全领域和产业势必会衍生出更为丰富、更为智能的应用实践。

当然，人工智能是一把双刃剑。做好人工智能自身安全是确保人工智能应用健康发展的根本保障。在人工智能快速发展的同时，应清醒地认识到需要产业各方以法律和伦理为界，确保人工智能技术安全可信赖，助力打造高度自治、智慧安全的网络空间。