第3章 无人系统感知安全

作为无人系统的基石，无人系统智能感知层可确保无人系统在复杂环境中准确、可靠地感知周围环境信息，并保障无人系统的正确决策和控制。随着5G通信网络和智能感知技术的兴起，以及传感器硬件设备的发展，无人系统的信息感知能力得到了显著提升，然而，随之而来的是智能感知系统面临着越来越多的安全风险和威胁。例如，腾讯科恩安全实验室利用对抗性样本攻击特斯拉公司的Model S所搭载的Autopilot自动驾驶系统中的视觉传感摄像头，令自动驾驶系统输出错误的识别结果，致使汽车驶入对向车道；360公司的独角兽研究团队通过GPS信号欺骗攻击使得大疆精灵3无人机错误认定了禁飞区，使得无人机无法起飞。因此，针对无人系统感知安全进行研究和学习，保障无人系统中感知系统的稳定可靠运行十分关键，通过学习本章的内容，读者将可以深入了解无人系统感知安全的重要性，掌握无人系统感知安全中相关攻击的基本原理和防御方法，为进一步研究和应用无人系统感知安全技术奠定基础。

本章要点

- 无人系统中的传感器介绍及相关的攻击原理与防御方法。
- 无人系统中针对感知数据的投毒攻击和对抗性样本攻击的基本原理与防御方法。
- 无人系统中的多源感知数据融合安全。

3.1 无人系统感知安全现状概述

无人系统的感知能力是指通过各种传感器（如激光雷达、摄像头、超声波传感器）获取环境信息的能力，是实现自主或远程控制的核心技术。随着5G通信网络和智能感知技术的兴起，以及传感器硬件设备的发展，无人系统的信息感知能力得到了显著提升，然而，随之而来的是其智能感知系统面临着越来越多的安全风险和威胁。

一方面，无人系统的感知系统可能会受到复杂自然环境的影响，例如极端高温和低温环境、高湿度环境、雨雪或雾霾环境、电磁干扰等，这些因素可能导致感知系统遭受扰动，以致传感器无法正常工作。另一方面，无人系统的感知系统也可能遭受到恶意攻击者的人为攻击，例如恶意干扰、欺骗攻击、物理破坏等，这些攻击可能导致感知信息对无人系统的决策系统造成误导，从而影响无人系统的正常运行，造成严重的安全后果。

此外，为了有效地提高无人系统的感知系统的鲁棒性，通常采用多传感器对多源数据进行信息感知，再通过数据融合技术将多源数据进行融合以进行进一步的分析和决策。然而，对于多源数据的感知和融合也使得数据收集、分析和决策过程中要面临更复杂的数据攻击，由于无人系统具有自主控制的特性，其对于此类攻击的防御能力更为欠缺。

在探讨无人系统感知安全方面面临的挑战时，可以将安全攻击大致划分为两大维度：一是针对无人系统传感器实施的攻击；二是针对传感器数据层的攻击。首先，对于传感器层面的攻击，其核心在于通过干扰、欺骗、物理损坏以及拒绝服务等方式，直接作用于无人系统搭载的各种传感器装置。例如，在自动驾驶汽车场景下，攻击者可能会干扰其激光雷达、摄像头或是GPS定位系统的信号，而在无人机集群操控中，则可能出现对超声波陀螺仪的干扰行为，这类攻击使得无人系统的关键传感器功能受损，无法准确且有效地捕获并反馈环境信息至决策中枢。其次，数据层面上的安全威胁则表现为对抗性样本攻击和投毒攻击两种形式。对抗性样本攻击是指攻击者巧妙地在原始数据上添加微乎其微却极具针对性的噪声扰动，这种细微改动足以诱使无人系统的决策模型在处理此类数据时，得出与预期结果相悖的输出结果，例如，攻击者可能对自动驾驶汽车摄像头捕捉到的交通标识进行难以察觉的人工篡改，经过篡改的交通标识能够成功欺骗车辆的感知系统，导致车辆误判并做出错误的行驶决策；投毒攻击则通常在无人系统决策模型的训练阶段悄然发生，攻击者蓄意篡改或操纵用于模型训练的传感数据及其标签，最终致使决策模型在实际应用中遇到特定输入数据时，产生错误的响应结果，比如，攻击者可能向无人机使用的地理信息数据集中注入“毒数据”，由此引发无人机在执行特定区域任务时出现严重的位置判断偏差，进而影响任务的完成度。此外，当无人系统依赖多源数据融合技术时，由于数据流经多个源头和使用者，在存储、传输和使用过程中面临的风险不容忽视，包括但不限于未经授权的访问、窃取和篡改等安全问题。以自动驾驶汽车为例，其运行过程中需整合雷达、激光雷达、摄像头等多种传感器及GPS定位数据，一旦这些数据链条中的任何一环遭受到攻击者的截取、伪造或篡改，就可能导致自动驾驶汽车出现偏离预定路线、意外停车、违反交通法规等严重影响安全的行为。因此，全方位保障无人系统感知层面的安全，对于确保其稳定、可靠运行至关重要。图3-1展示了无人系统中的感知攻击分类。

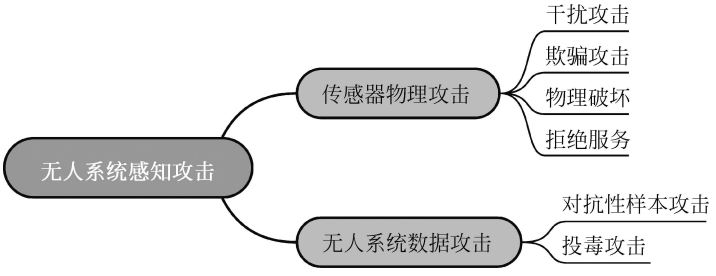

图 3-1　无人系统中的感知攻击分类

在构建无人系统感知安全的坚固防线时，除了传统意义上的加密、数字签名、安全审计等措施确保数据的完整性与机密性外，还可以从数据和模型两个核心方向来提高系统对感知攻击的抵抗力。从数据层面着手，可以运用先进的异常检测算法来甄别潜在的对抗性样本和投毒攻击，通过监测输入数据的特征变化和统计规律，从而发现那些有意为之的、旨在误导系统的恶意数据。同时，引入异常数据转化技术，对识别出的对抗性样本进行预处理，

还原其原有属性，使之在输入至决策模型时不会引起误判。此外，充分利用多传感器融合的优势，通过各传感器间的互补性和冗余性，即便某一传感器受到攻击或失效，其他传感器仍能保证整体系统的稳定感知，降低单点故障的风险。在模型层面，现代防御技术涵盖了对抗性训练、防御性蒸馏、多模型集成和鲁棒性正则化等多个领域。对抗性训练通过模拟攻击环境训练模型，使其在面对真实攻击时能够保持稳定性能；防御性蒸馏则是通过知识蒸馏的方法提炼模型的核心知识，减轻对抗样本的影响；多模型聚合则是通过多种模型共同投票或平均预测，分散单一模型可能存在的弱点；鲁棒性正则化则在训练过程中引入额外约束，鼓励模型学习对输入扰动的不变性，从而增强模型本身的稳健性。

至于传感器硬件本身，通过物理手段，如采用电磁屏蔽、防水防尘结构和专用滤波技术，可以显著减少外部环境因素对传感器信号质量的影响，并提高其抵抗各种干扰信号的能力，确保无人系统在复杂环境下的稳定感知效能。

接下来，本章将分节详细介绍传感器相关攻击、对抗性样本攻击、投毒攻击的基本原理、技术细节及相关防御方法。此外，本章还将探讨多源数据融合的安全问题，并针对基于联邦学习的无人系统中的拜占庭攻击进行案例分析。

3.2　传感器攻击及防御

在本节中，我们将介绍无人系统中常见的传感器及其面临的直接攻击和安全风险。具体而言，针对无人系统的传感器攻击主要分为两类：干扰攻击和欺骗攻击。

- 干扰攻击：这类攻击的目的是使传感器失效或降低其性能，这可以通过多种手段实现，例如物理破坏、电磁干扰，或者通过污染环境（如在摄像头前投射强光或烟雾）来影响传感器的准确性。干扰攻击的关键特点是它们直接影响传感器的物理或电子功能，从而使其无法准确收集数据。
- 欺骗攻击：这类攻击的目的是操纵传感器所接收的数据，从而使系统做出错误的判断或反应，例如，攻击者可能通过向雷达或声呐系统发送伪造的信号，导致无人系统错误地识别出虚假目标或障碍。欺骗攻击并不直接破坏传感器的功能，而是通过操纵数据误导系统。

接下来，本章将从具体的传感器出发，介绍其相关的攻击类型及相应的防御措施，图3-2展示了自动驾驶系统中的常见传感器。

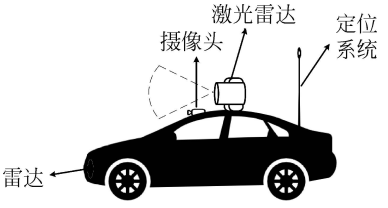

图 3-2　自动驾驶传感器

3.2.1　激光雷达

激光雷达(LiDAR)是一种光学传感器，它通过向目标照射一束脉冲激光，并测量发射与接收脉冲信号的时间间隔，来计算与目标物体的距离。在无人系统中，激光雷达可用于提供高精度、高稳定性的三维数据。例如，在自动驾驶系统中，激光雷达可以用于实时感知车辆周围的环境，从而帮助自动驾驶系统进行高效路径规划并做出驾驶决策。

当前，针对激光雷达的攻击主要集中于激光雷达干扰和激光雷达欺骗两类。激光雷达干

扰攻击是一种拒绝服务攻击，攻击者通过发送与激光雷达传感器相同波长但强度更高的光，使得传感器无法获取有效的光波。研究证明，利用与激光雷达相同波长的强光源“致盲”激光雷达，可使激光雷达无法在光源方向上感知目标[6]。激光雷达欺骗攻击的一类典型攻击在于攻击者可以录制来自于激光雷达传感器的合法信号，并将这些信号中继到同一辆自动驾驶汽车的另一个激光雷达传感器，从而导致无人系统对于感知对象的位置相比实际的位置更近或更远；另一种典型攻击是攻击者通过伪造激光雷达传感器信号来模拟一个对象，并将伪造信号注入激光雷达传感器，可能会导致自动驾驶车辆认为它正在接近一个大物体并进行紧急制动[7]。

防御者可以通过引入数据采集的随机性来防御激光雷达欺骗攻击，当激光雷达的探测时间被设置为随机时，攻击者将难以准确获取激光雷达接收信号的具体探测时间窗口，从而难以成功发送伪造信号，进而抵御欺骗攻击。同时，融合来自不同传感器的数据能够有效地稳定智能感知层的性能以缓解欺骗攻击。

3.2.2 雷达

无人系统中的雷达可分为毫米波雷达和超声波雷达，其工作原理与激光雷达类似，均通过发射和接收电磁波或超声波来计算目标物体的距离和速度。相比于激光雷达，这些雷达系统的检测范围更广，探测距离更远，且对天气条件的依赖性较低。

针对雷达的攻击与激光雷达原理基本相同，可分为雷达干扰攻击和雷达欺骗攻击。雷达干扰攻击主要是利用干扰器或信号发生器来干扰雷达感知系统。2016 年的 DEFCON 大会上有研究团队提出利用信号发生器和频率倍增器生成电磁波攻击特斯拉的 Autopilot 自动驾驶系统，自动驾驶系统在该过程中受到了显著损伤[8]，该团队同样利用超声波干扰器攻击了四辆汽车的停车辅助系统，结果表明在干扰攻击下车辆无法有效检测周围的障碍物。雷达欺骗攻击同样是通过伪造信号或是中继合法信号来实现雷达欺骗。通过实施信号中继雷达欺骗攻击，一个 121 米远的物体在雷达设备的感知中仅显示为 15m 远[9]。

针对雷达攻击的防御方法也与激光雷达相似，利用多传感器的信息冗余来增强感知系统的鲁棒性，或是将探测时间窗口设定为随机以防御欺骗攻击。一种新颖的防御方法是物理挑战-响应身份验证 (PyCRA)，该方法通过发送随机挑战信号来检查周围环境，并通过假设性检验来确定恶意信号是否高于噪声来检测恶意信号，能够有效地防御欺骗攻击[10]。

3.2.3 摄像头

随着人工智能技术的不断发展，基于图像的目标检测和识别技术已有了显著进步，摄像头成为无人系统中应用和研究最广泛的传感器，目前无人系统搭载的视觉感知算法能够有效地帮助其进行决策，执行不同的任务。在典型的自动驾驶系统中，摄像头被用于车辆、行人、车道检测，以及红绿灯和交通标志的识别，从而在自动驾驶环境感知的各方面发挥作用。然而，作为被动型感光设备，摄像头受环境影响较大，较容易受到干扰。

针对摄像头的直接攻击可大致分为摄像头干扰攻击和摄像头欺骗攻击。摄像头干扰攻击是攻击者利用强光使得摄像头暂时“致盲”甚至永久“失明”。BlackHat 2015 大会上研究者向摄像头发射强光，致使摄像头的自动曝光功能无法工作，导致捕获到的图像过度曝光，无法被神经网络模型正常识别，从而导致无人系统无法正常决策[11]。摄像头欺骗攻击指攻

击者通过模拟虚假光学信号，致使摄像头采集到虚假的图像信息。例如，通过投影仪改变地面平面的外观，可影响光流摄像头采集的信息，攻击者可以借此控制无人机[12]。

当前针对摄像头干扰攻击的防御可通过物理层面增加摄像头的数量，或者在摄像头中集成可拆卸的近红外光滤波器，来减轻或消除摄像头被“致盲”的风险[11]。预测性防御方法利用预测分析来预测摄像头捕获的未来帧图像，并将未来帧图像与感知到的图像进行比较以防御干扰攻击，针对部分摄像头欺骗攻击，该预测性方法也可以进行有效防御[13]。

针对摄像头的对抗性样本攻击和数据投毒攻击是另一类数据层面的广泛感知攻击，将在本章的3.4节和3.5节进行详细讨论。

3.2.4　定位系统

全球卫星定位系统（Global Positioning System，GPS）和北斗卫星定位系统通过卫星信号确定当前的经纬度和海拔，是目前应用最广泛的定位手段。在无人系统中，卫星定位系统通常用于获得无人设备的位置。

卫星定位系统同样会受到干扰攻击和欺骗攻击的影响，干扰攻击指卫星定位系统遭受到干扰设备的攻击，大量的强信号噪声导致GPS定位系统无法正常工作，其本质上是一种拒绝服务攻击；GPS欺骗攻击是通过伪造虚假GPS定位信号或者重放GPS定位信号，使得GPS接收传感器解算出错误的位置和时间信息，前者被称为生成式GPS欺骗，后者被称为转发式GPS欺骗。图3-3展示了一种典型的针对无人机网络的GPS欺骗攻击，其中左半部分表示正常的GPS操作，右半部分表示GPS欺骗攻击。首先，攻击者跟踪目标无人机的GPS信号，并伪造一个功率比合法GPS信号更强的信号；接着，由于合法的GPS信号被伪造的GPS信号压制，受害无人机接收到伪造的GPS信号并导致导航信息错误。美国得克萨斯大学的研究团队在2013年利用GPS欺骗攻击使得地中海上的一个游艇偏离航向[14]；DEFCON 2015大会上，奇虎360的独角兽研究团队通过GPS信号欺骗攻击对大疆无人机进行了禁飞区欺骗，通过将禁飞区内的GPS信号传递给大疆无人机的GPS定位接收传感器，使得无人机无法正常起飞。

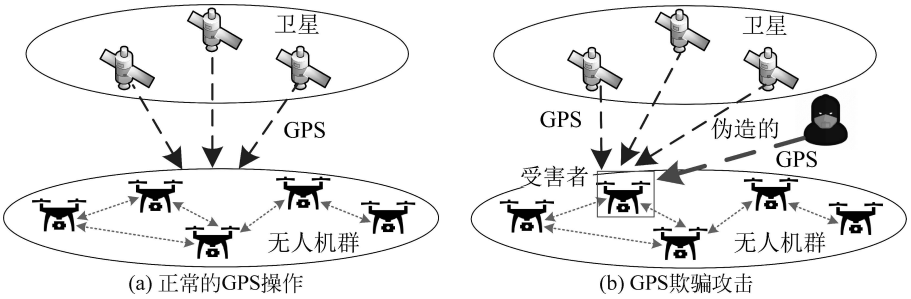

图 3-3　典型的针对无人机网络的GPS欺骗攻击

针对卫星定位系统攻击的通用防御方法包括监测卫星定位信号的绝对强度和相对强度来进行攻击检测、检查卫星定位信号之间的时间间隔或时间比较以检测卫星定位信号的真实性、通过监控卫星识别代码和接收到的卫星信号数量来判断卫星定位信号的来源；尽管这些方法简单且成本低廉，足以防御简单的卫星定位欺骗攻击，但对于更复杂的攻击形式，它们可能不够有效。学术界已提出多种策略来防御卫星定位欺骗攻击，这些策略利用了卫

星定位信号的物理层特性、其他导航技术及密码学方法。每个卫星的信号在物理层面上都具有独有的特征，这是因为接收器与各个卫星之间的距离和相对位置存在差异。相比之下，人为制造的GPS干扰信号通常通过特定的发射设备产生，使这类欺骗性信号在物理层面上表现出相似或一致的特性。

3.2.5 惯性测量单元

惯性测量单元（Inertial Measurement Unit，IMU）是一个集成了多个传感器的装置，通常由陀螺仪和加速度计两类传感器构成，可以用于测量物体的三轴姿态角及加速度；同时，IMU可以辅助GPS定位系统实现可靠的定位。IMU在无人系统中被广泛应用于准确的动作和姿态测量，例如，在无人机系统中，通过IMU中的微机电系统陀螺仪和加速度计，可以对无人机飞行中的加速度与角速度进行测量，从而实现对无人机飞行状态的稳定控制。

针对惯性测量单元的攻击主要是对于陀螺仪的干扰攻击，在无人机系统中，微机电系统陀螺仪被用来感知角速度，但它会受到共振频率的影响。外部干扰的振动频率与陀螺仪的固有频率一致时，会引发共振现象，进而影响陀螺仪的准确性，甚至造成传感器损坏。通过产生与无人机内置陀螺仪相同振动频率的超声波噪声来干扰无人机，可以导致陀螺仪失灵，使得惯性测量单元无法正常运作，从而导致无人机失去对姿态的控制，最终导致无法稳定飞行并坠毁[15]。

对于此类干扰攻击，最简单的防御方法是进行物理隔离，例如可以使用泡沫和镍纤维保护罩来实现物理隔离保护方案，减弱干扰信号对于微机电系统陀螺仪的影响[16-17]。

3.3 对抗性样本攻击及防御

对抗性样本攻击基于机器学习模型对于输入数据的敏感性，通过对抗性构造策略针对原始数据进行人眼几乎不可察觉的微不足道的修改，精心构造出对抗性样本，使得机器学习模型产生错误的输出。深度学习模型对于输入数据的敏感性更高，更易受到对抗性样本攻击的欺骗。在无人系统中，由于深度学习模型在智能感知层和自主决策层都被广泛使用，因而对抗性样本攻击对无人系统构成了相当严重的威胁。

在本节中，首先介绍对抗性样本攻击的基本定义、原理及相关概念。随后，讨论对抗性样本攻击在无人系统中所造成的风险和研究进展。最后，讨论无人系统中针对对抗性样本可能的防御方法和研究方向。

3.3.1 对抗性样本攻击基本原理

1. 对抗性样本攻击定义

首先，本节给出对抗性样本的基本定义：考虑一个数据点$x \in \mathbb{R}^d$属于一个具体的类$\mathcal{C}_i$，对抗性样本即通过对该样本点的属性进行微调或扰动，使其成为一个新的样本点$\hat{x}$，且该样本点$\hat{x}$会被目标分类器f误分类为$\mathcal{C}_t$，$\hat{x}$就被称为一个对抗性样本。生成并利用对抗性样本，实现使目标分类器f误分类并达到攻击者目的的过程被称为对抗性样本攻击。例如，如图3-4所示，原始图像是一个禁止输入标志的图像，通过对抗性样本生成添加噪声，得到新的

对抗性攻击图像，攻击图像将被目标分类器分类成限速标志。在某些场景下，攻击者的目标可能不是令分类器 f 将 $\hat{x}$ 分类到一个具体的目标类 $\mathcal{C}_t$，而是只需要令 $\hat{x}$ 偏离 x 的类别 $\mathcal{C}_i$，这种攻击也是一种对抗性样本攻击，被称为无目标对抗性样本攻击，相应的，将对抗性样本误分类为特定的分类被称为有目标对抗性样本攻击。本节主要从有目标对抗性样本攻击出发，给出对抗性样本攻击的形式化定义。

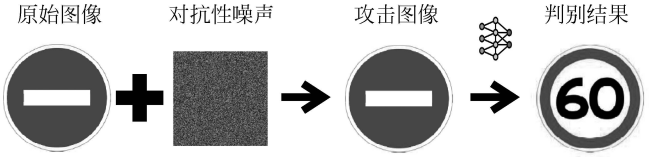

图 3-4　对抗性样本攻击示意

假定存在一个映射函数 $\mathcal{A}:\mathbb{R}^d \to \mathbb{R}^d$，$\mathcal{A}(x)=\hat{x}$，即 $\mathcal{A}(x)$ 是扰动后的目标样本，给出有目标对抗性样本攻击的基本定义如下。

定义 3.1 (有目标对抗性样本攻击)　令 $x \in \mathbb{R}^d$ 是一个属于 $\mathcal{C}_i$ 的数据点。定义一个目标类别 $\mathcal{C}_t$，有目标对抗性样本攻击可被定义为一个映射函数 $\mathcal{A}:\mathbb{R}^d \to \mathbb{R}^d$ 使得 $\mathcal{A}(x)=\hat{x}$，且针对目标分类器 f 令 $f(x)=\mathcal{C}_i$ 且 $f(\hat{x})=\mathcal{C}_t$。

接着，本节将介绍最通常使用且最简单的对抗性样本攻击——加性对抗性样本攻击，假定映射函数 $\mathcal{A}$ 是一个线性函数，并且在原始输入上添加扰动 p。

定义 3.2 (加性对抗性样本攻击)　令 $x \in \mathbb{R}^d$ 表示一个属于 $\mathcal{C}_i$ 的数据点。定义一个目标类别 $\mathcal{C}_t$，加性对抗性样本攻击通过在原始输入上添加扰动 $p \in \mathbb{R}^d$ 使得 $\hat{x}=x+p$，且针对目标分类器 f 令 $f(x)=\mathcal{C}_i$ 且 $f(\hat{x})=\mathcal{C}_t$。

加性对抗性样本攻击具有两大优势：其一，加性对抗性样本攻击保证了输入空间保持不变，例如，原始输入 x 是 $\mathbb{R}^d$ 上的一张图片，那么对抗性样本 $\hat{x}$ 也保持在 $\mathbb{R}^d$ 内，并仍然能表示成一张图片，通用的映射函数可能不能保持这种关系（子采样函数或者提取特征的密集矩阵）；其二，加性攻击允许用简单的几何图形进行可解释的分析，而通用的映射函数分析起来可能相当复杂。

为了更好地理解对抗性样本攻击，本节将从机器学习分类器的角度出发针对对抗性样本攻击进行分析与形式化描述。考虑一个常见的多分类场景，即存在类 $\mathcal{C}_1,\cdots,\mathcal{C}_k$，该 k 个类的决策边界可由 k 个判别函数 $g_i(\cdot)$ 表示，$i=1,\cdots,k$。如果要将对抗性样本 $\hat{x}$ 分类为目标类 $\mathcal{C}_t$，这些判别函数需要满足如下条件：

$$g_t(\hat{x}) \geqslant g_j(\hat{x}),\ \text{对所有}\ j \neq i \tag{3.1}$$

即目标类别 $\mathcal{C}_t$ 应该相比于分类器可分的其他类别有更大的判别值 $g_t(\hat{x})$。重写式 (3.1)，可得 $g_t(\hat{x}) \geqslant \max\limits_{j\neq i}\{g_j(\hat{x})\}$。因此，对抗性样本攻击的目标是找到一个 $\hat{x}$ 以满足式 (3.1)。事实上，在具体实现对抗性样本攻击的过程中，可能存在一组 $\hat{x}$ 都能满足式 (3.1)，这些样本都允许被作为潜在的攻击样本。然而，有些攻击样本相比于其他攻击样本更接近原始输入 x，那么，为了选择最优的攻击，往往需要通过对扰动幅度添加额外的标准，即希望所添加的扰动尽可能小或者尽可能使扰动后生成的对抗性样本更接近原始输入。将这些标准添加到问题中，会得到一个形式化的优化问题，根据该优化问题，给出如下最常用的三类对抗性样本攻击定义。

定义 3.3 (最小范数对抗性样本攻击) 最小范数对抗性样本攻击通过解如下的优化问题，得到最优的扰动后对抗性样本 $\hat{x}$：

$$\begin{aligned} \underset{\hat{x}}{\text{minimize}} \quad & \|\hat{x}-x\| \\ \text{s.t.} \quad & \max_{j\neq t}\{g_j(\hat{x})\}-g_t(\hat{x})\leqslant 0 \end{aligned} \tag{3.2}$$

$\|\cdot\|$ 可以是用户定义的任意范数。

最小范数对抗性样本攻击的目标是在最小化扰动的量级的同时，确保对抗性样本被分类器误分类到指定的目标类别，如果约束集为空，则意味着对目标分类器的原始输入无法进行成功的攻击。

最小范数对抗性攻击的另一种替代形式是最大允许对抗性样本攻击，定义如下。

定义 3.4 (最大允许对抗性样本攻击) 最大允许对抗性样本攻击通过解如下的优化问题，得到最优的扰动后对抗性样本 $\hat{x}$：

$$\begin{aligned} \underset{\hat{x}}{\text{minimize}} \quad & \max_{j\neq t}\{g_j(\hat{x})\}-g_t(\hat{x}) \\ \text{s.t.} \quad & \|\hat{x}-x\|\leqslant \tilde{p} \end{aligned} \tag{3.3}$$

$\|\cdot\|$ 可以是用户定义的任意范数，$\tilde{p}>0$ 表示攻击的程度。

在最大允许对抗性样本攻击中，攻击的大小以 $\tilde{p}$ 为界限，目标函数是找到尽可能使 $\max_{j\neq i}\{g_j(\hat{x})\}-g_t(\hat{x})$ 为负的 $\hat{x}$。如果最优的 $\hat{x}$ 仍无法使目标值为负，则不可能对目标分类器的原始输入进行成功攻击。

第三类形式化的对抗性样本攻击定义是一个基于正则项的无约束优化问题。

定义 3.5 (正则化对抗性样本攻击) 正则化对抗性样本攻击是指通过解如下的优化问题，得到最优的扰动后对抗性样本 $\hat{x}$：

$$\underset{\hat{x}}{\text{minimize}} \quad \|\hat{x}-x\|+\lambda(\max_{j\neq t}\{g_j(\hat{x})\}-g_t(\hat{x})) \tag{3.4}$$

$\|\cdot\|$ 可以是用户定义的任意范数，正则化参数 $\lambda>0$。

正则化对抗性样本攻击中的参数被称为拉格朗日乘子，直观来看，正则化对抗性样本攻击试图同时最小化两个冲突的目标，依赖于 λ 的选择，可以控制两个优化目标的相对重要程度。

事实上，三类攻击在选择合适的参数 $\tilde{p}$ 和 λ 时是等效的，具有相同的最优解。从几何意义上来讲，对抗性样本攻击就是使得扰动后的对抗性样本的 $\hat{x}$ 处于目标分类 $\mathcal{C}_t$ 的决策边界内。

2. 对抗性样本攻击分类

表3-1展示了基本的对抗性样本攻击分类情况。根据对抗性样本攻击的攻击目标进行分类，对抗性样本攻击可以分为两类：有目标对抗性样本攻击及无目标对抗性样本攻击。有目标对抗性样本攻击前文已经进行了详尽的介绍；无目标对抗性样本攻击与有目标对抗性样本攻击的攻击目标不同，无目标对抗性样本仅需将扰动后的对抗性样本偏离原始输出的类别，例如，原始输入 $x\in\mathcal{C}_i$，无目标对抗性样本的目标是使得扰动后样本 $\hat{x}\notin\mathcal{C}_i$，相比于有目标对抗性样本攻击的攻击条件更为松弛。无目标对抗性样本攻击的约束集可以表示为

$$\Omega=\{\hat{x}|g_i(\hat{x})-\min_{j\neq i}\{g_j(\hat{x})\}\leqslant 0\} \tag{3.5}$$

表 3-1　对抗性样本攻击分类

分 类 依 据	分 类 结 果
攻击目标	有目标对抗性样本攻击；无目标对抗性样本攻击
攻击方法	加性对抗性样本攻击；最小范数对抗性样本攻击；正则化对抗性样本攻击
攻击者能力	黑盒对抗性样本攻击；白盒对抗性样本攻击

根据对抗性样本攻击者的能力进行分类，对抗性样本攻击可以分为两类：白盒对抗性样本攻击及黑盒对抗性样本攻击。白盒对抗性样本攻击假定攻击者具有对于分类器的全面知识，即攻击者针对每个类别都知道具体的判别函数 $g_i(\cdot)$。对于线性分类器而言，这要求攻击者知道所有的决策边界；对于深度神经网络而言，这要求攻击者知道具体的神经网络参数值和网络架构。因此，白盒对抗性样本攻击被认为是攻击者能够设计的最优对抗性样本攻击。黑盒对抗性样本攻击意味着攻击者无法知道分类器的内部知识，只能够通过随机添加噪声来实现对抗样本的生成；此外，攻击者对一个黑盒分类器的访问次数通常被限制在一个固定的数值。相比于白盒对抗性样本攻击，黑盒对抗性样本攻击通常更加具有挑战性，也更符合实际场景。

3. 典型对抗性样本攻击

- 随机噪声攻击：随机噪声攻击是最简单的一类对抗性样本攻击，是无目标黑盒对抗性样本攻击的代表，通过在原始样本上添加不包含任何信息的噪声扰动 (例如，高斯随机噪声)，使得扰动后样本偏移原始的分类。然而，这类攻击很难奏效，扰动后的样本容易和原始样本区分，甚至人眼也能进行分辨。
- 语义攻击：语义攻击也是一类图像识别领域典型的无目标黑盒对抗性样本攻击，通过反转图片的所有像素强度 (例如，改变所有像素的符号) 来得到一个负面图像，使得分类器对于该图像进行误分类。然而，这类攻击同样难以奏效，扰动后的样本同样很容易进行区分[18]。
- 快速梯度符号法：快速梯度符号法（Fast Gradient Sign Method，FSGM）是最著名的无目标白盒对抗性样本攻击之一，并且能够直接推至广有目标攻击[19]。快速梯度符号法的基本原理是在图像的每个像素处沿着梯度符号的方向进行梯度更新，其生成的对抗性扰动可以表示为

$$p = \varepsilon \text{sign}(\nabla_x J(\theta, x, y)) \tag{3.6}$$

  其中，ε 表示扰动的具体大小，该扰动可以通过反向传播来进行计算，生成的对抗性样本 $\hat{x}$ 为 $\hat{x} = x + p$。FSGM 能够有效地生成对抗性样本的原理在于，深度神经网络同样具有先行性质，高维空间中的线性行为足以引起对抗性样本，在 x 上加上计算得到的梯度方向，使得修改后的样本经过分类网络时的损失值比修改前的样本经过分类网络时的损失值更大，因而更容易导致误分类。例如，快速梯度符号法通过生成对抗性图像样本，会使得分类器将熊猫分类为长臂猿。
- 基本迭代法：基本迭代法（Basic Iterative Methods，BIM）是快速梯度符号法的一种变体，它通过在多次迭代中重复性地在样本 x 上添加更精细的扰动，来实现目标分类器对于扰动后对抗性样本的误分类[20]。具体而言，在每次迭代中，BIM 通过限定属性变化范围来避免单个属性发生较大的变化，以图像为例，可通过裁剪像素值来避

免每个像素发生较大的变化。

- 投影梯度下降法：投影梯度下降法（Projected Gradient Descent，PGD）同样是一种迭代攻击，可视为BIM和FSGM的一种拓展[21]。在每次迭代扰动过程后，通过投影函数Π将对抗性样本投影到原始输入x的ε球体中，即每个像素最大被允许的变化范围为ε：

$$\hat{x}^T = \Pi_\varepsilon(x^{T-1} + \gamma \text{sign}(\nabla_x J(\theta, x, y))) \tag{3.7}$$

其中，$\hat{x}^T$是第T轮迭代后的对抗性样本，γ是每轮迭代过程中添加的扰动大小。与BIM不同，PGD对于x进行随机初始化，即添加一个在$(-\varepsilon, \varepsilon)$范围内的随机噪声。PGD被视为最难以防御的基于梯度的对抗性样本攻击，PGD可以简单地通过调整迭代过程来将无目标对抗性样本攻击转化为针对类别$\mathcal{C}_T$的有目标对抗性样本攻击：

$$\hat{x}^T = \Pi_\varepsilon(x^{T-1} - \gamma \text{sign}(\nabla_x J(\theta, x, t))) \tag{3.8}$$

- Deepfool攻击：Deepfool攻击通过寻找原始输入到对抗性样本决策边界的最近距离来生成对抗性样本，并通过寻找到分类器的超平面来最终使得分类器误分类对抗性样本[22]。相比于FSGM，Deepfool引入了更小的扰动。
- 单像素攻击：单像素攻击通过修改一个像素来生成对抗性样本，以降低防御者对于对抗性样本的检测能力[23]，该攻击通过应用差分进化算法来寻找修改像素的最优解。
- Carlini-Wagner攻击：Carlini和Wagner实现了一种基于正则项的有目标对抗性样本攻击，能够通过现有大多数的对抗检测防御[24]。Carlini和Wagner针对对抗性样本攻击重新定义了其优化问题：

$$\begin{aligned} \underset{\hat{x}}{\text{minimize}} \quad & \|p\|_1 + \lambda g(x+p) \\ \text{s.t. } & x + \|p\|_1 \in [0,1]^n \end{aligned} \tag{3.9}$$

- 基于生成对抗网络（Generative Adversarial Network，GAN）的攻击：通过GAN生成对于人类而言看起来更加自然的对抗性样本，通过最小化输入样本之间的内部表示的距离，结合对抗性样本的目标函数，生成对抗性样本[25]。由于基于对抗性样本的攻击不需要原始神经网络梯度，因此是一种黑盒对抗性样本攻击。

3.3.2 对抗性样本攻击防御方法

针对对抗性样本攻击的防御方法可大致分为三类[26]，表3-2展示了具体的防御方法与主要技术。

表 3-2 对抗性样本攻击防御方法及其主要技术

防御方法	主要技术
梯度掩蔽/模糊	防御性蒸馏；碎片化梯度；随机梯度；梯度爆炸和消失
对抗性防御	正则化方法；对抗性重训练
对抗性样本检测	辅助模型分类对抗性样本；统计学检测对抗性样本；检查预测一致性

- 梯度掩蔽/模糊：由于大多数白盒对抗性样本攻击都需要基于分类器的梯度信息进行攻击，因此对梯度进行掩蔽或者隐藏能够误导攻击者，从而防御对抗性样本攻击。
- 对抗性防御：通过在训练集中主动添加部分对抗性样本（具有正确的标签），并且对

深度神经网络分类器参数进行重新训练可以有效增强其鲁棒性，从而正确分类对抗性样本。

- 对抗性样本检测：通过研究自然样本或者说良性样本的数据分布，在将样本输入进机器学习模型之前首先对其进行对抗性样本检测并禁止对抗性样本输入目标分类器。

1. 梯度掩蔽/模糊

由于大多数攻击者的策略依赖于获取目标分类器的梯度信息，所以通过掩蔽或模糊梯度可以有效隐藏这些信息，从而有效防御基于梯度的对抗性样本攻击。

- 防御性蒸馏：知识蒸馏利用大尺寸模型的输出作为标签去训练一个小尺寸神经网络，能够减小深度神经网络架构的大小。防御性蒸馏重构了蒸馏过程，通过在训练过程中的温度调整，使得目标输出类的分数接近于1，而所有其他类的分数接近于0，在计算机中，其对应的梯度也由于浮点数机制而接近于0，从而抑制了FSGM、Deepfool等基于梯度的攻击[27]。
- 碎片化梯度：通过对输入样本进行预处理来实现针对对抗性样本攻击的防御，其通过添加一个非平滑或者不可微的预处理器，随后再进行深度神经网络模型的训练[28]。由于在参数上不可微，因此训练后的分类器对于参数的微小变化不敏感，从而能够防御基于梯度的对抗性攻击。例如，温度计编码通过一个预处理器使得图像的像素离散化，成为一个独热编码，从而通过这些编码来进行神经网络训练。
- 随机梯度：随机梯度防御方法通过使深度神经网络具有随机性来迷惑对抗性攻击者[29]。例如，在进行样本分类预测时，随机在深度学习模型中的每一层丢弃掉一些神经元，从而来防御对抗性样本攻击；或通过训练一组分类器，并在对数据进行评估时随机选择其中一个分类器进行预测，可以降低攻击者的准确率。由于攻击者无法确定实际使用的分类器，因此增加了攻击的难度。
- 梯度爆炸和消失：利用生成模型将可能的对抗性样本投影到良性数据流形上，再作为样本输入到目标模型中，这些生成模型可以被视为一种净化器，将对抗性样本转换为良性样本[30-31]。该防御通过令模型每一层的偏导数累积乘积，使得梯度变得极小或不规则地变大，进而使对抗性样本生成变得困难。

2. 对抗性防御

对抗性样本可能威胁任何机器学习算法的分类准确性，但相应地，基于对抗性样本的防御技术也能够帮助构建更精确、更鲁棒的模型。这类防御的目标是不仅在自然样本上实现高准确率，也要确保对抗性样本的高准确预测。事实上，对抗性防御与两方攻防博弈的过程非常相似：攻击者试图通过在原始样本上添加微小的扰动生成能够误导分类器的对抗性样本；而防御者则致力于开发能够精确识别这些样本的鲁棒模型，使得攻击者难以找到能够欺骗这些模型的对抗性样本。

- 正则化方法：为了增强机器学习模型的稳定性，可以在模型参数上添加正则项。例如可通过限制莱布尼茨常数来增加模型输出的稳定性[32]，可引入深度收缩网络来规范训练，这种网络在训练过程中通过在反向传播框架的每一层添加惩罚项来减少输入样本变化对输出的影响，从而降低每一层输出的变化幅度[33]。
- 对抗性重训练：通过在训练集中加入生成的对抗性样本，并使用正确的标签对其进行标记和重新训练，能够有效提升神经网络模型对于对抗性样本攻击的鲁棒性，对抗性

训练可以针对FSGM、PGD等攻击实现有效防御[19]。

3. 对抗性样本检测

检测对抗性样本是另一种主要的分类器保护方法，这种方法不会直接对模型输入进行预测，而是先判断输入样本是良性的还是对抗性的。如果输入被识别为对抗性样本，分类器则会拒绝给出预测标签。有效的对抗性样本检测方法应当能够应对包括零知识对手、完全知识对手和有限知识对手在内的各种威胁模型，同时确保能正确识别对抗性样本，并且尽可能减少对良性样本的误判。

- 辅助模型分类对抗性样本：通过设计辅助模型来区分对抗性和良性样本，例如在训练深度学习模型时添加一个专门的对抗性样本类别[34]，用以识别对抗性样本。
- 统计学检测对抗性样本：通过探索对抗性和良性样本在统计特性上的差异，例如使用主成分分析或最大均值差异测试[34-35]，来检测对抗性样本。
- 检查预测一致性：通过微调模型参数和输入样本来测试预测的一致性，正常样本通常显示出稳定的预测结果，而对抗性样本则显示出较大的预测差异。例如，使用Dropout技术来随机化分类器，检测统一样本在随机化后是否显示出显著的预测差异，以判断其是否为对抗性样本[36]。

3.3.3 无人系统中的对抗性样本攻击

对抗性样本攻击在无人系统中通常被视为黑盒攻击，这是因为攻击者难以获得系统内部的神经网络模型的详细参数和结构。在实际应用中，这类攻击主要在物理环境中进行，攻击者通过对现实世界中的物体进行细微而不易察觉的修改，使传感器在不同的角度、距离和光照条件下感知到扭曲的信息。这样的干预使得传感器采集的数据变成对抗性样本，进而误导无人系统的感知模块。

针对自动驾驶系统的对抗性样本攻击通常着重于对物理世界中自动驾驶模型所感知的物体进行扰动使其成为对抗性样本，如广告牌、车道线、交通标志等。

- 广告牌误导：攻击者通过摆放可误导自动驾驶车辆摄像头的人工广告牌，能够在不同视角、距离、光照和天气条件下，使自动驾驶车辆的转向角度产生显著误差。攻击者还可生成与真实广告牌视觉上相似的对抗性广告牌，诱导车辆偏离预定路线。
- 车道线和交通标志篡改：攻击者可通过对抗性样本攻击方法，修改车道线，干扰自动驾驶系统的视觉感知，从而诱导车辆偏离预定行进路线。同时，攻击者可篡改交通标志为受扰动的交通标志，使得无人驾驶汽车误识别，造成严重的交通安全风险。
- 增强车辆隐形能力：攻击者可通过制作特殊的车辆纹理，使得目标车辆在无人驾驶系统中的车辆检测器中被忽视，降低其准确率，造成交通事故威胁。

除了针对摄像头的对抗性样本攻击之外，针对激光雷达的物体检测模型同样可以通过对抗性样本攻击来误导无人驾驶系统。攻击者可通过一种白盒优化的方法，生成对抗性的点集，并且通过激光将这些点注入障碍物的原始点云中，从而实现对抗性样本攻击，进而误导激光雷达的物体检测系统[7]，在百度Apollo发布的模拟器上使用激光雷达传感器数据的实验结果表明，该攻击的平均成功率高达90%。攻击者可针对激光雷达实施黑盒攻击，该方法通过将对抗性样本插入点云中来迷惑物体检测器[37]，相关数据集上进行的实验显示，该攻击的平均成功率为80%。

在无人机领域，对抗性样本攻击主要涉及以下两方面。

（1）视觉对抗性样本，即通过构造对抗性样本来干扰无人机搭载的摄像头等视觉传感器。这种攻击手段包括向物体添加纹理、引入遮挡、添加对抗性扰动贴纸等方式，旨在使无人机的目标检测、识别或跟踪模型产生错误的识别结果。

（2）信号对抗性样本，攻击者直接修改传感器接收的数据（如激光雷达点云，微波雷达信号等），通过微小的扰动使得目标跟踪和识别模型对目标进行误分类。

目前对无人机系统中的对抗性样本攻击研究仍处于起步阶段。文献[38]提出了两种基于前向倒数和优化的对抗性样本攻击方法，专注于针对无人机的导航系统构造对抗性样本。这些方法能够构造出与摄像机发送的原始图像几乎无法区分的对抗性样本，但却能显著改变无人机导航回归模型的输出，对无人机的导航和控制构成相当大的威胁。仍然存在对无人机的对抗性样本攻击缺乏合适防御方法和相应评估标准的问题，迫切需要解决。

3.4　投毒攻击及防御

投毒攻击与对抗性样本攻击有所不同，它通常发生在机器学习模型的训练阶段而非预测阶段，投毒攻击的目标是通过修改训练数据或者模型参数，从而影响模型在其原本任务上的性能或是向模型中注入特定的后门，在具有特定触发器的输入上产生错误的预测。在投毒攻击中，通常假设攻击者具有向训练集贡献数据的能力，或者能够直接控制训练数据本身。在无人系统中，倘若感知模型和决策模型被投毒攻击注入后门，有可能对无人系统造成误导，从而造成严重的安全风险。

3.4.1　投毒攻击基本原理

在机器学习领域，投毒攻击通常出现在有监督学习环境中，涉及数据收集、预处理、训练和测试的各个阶段。这类攻击主要通过在数据收集或预处理阶段注入恶意数据，导致模型在预测时表现异常。具体而言，投毒攻击定义为：攻击者将少量精心设计的恶意样本注入训练数据集，使得模型在训练过程中学习到这些样本的特征，从而在实际应用中引导模型产生错误的预测，破坏模型的可用性和完整性。在数据收集阶段，攻击者可能通过提供带毒的样本给机器学习服务提供商，或在数据标注过程中添加恶意标签；在预处理阶段，特权攻击者可以修改训练数据，包括添加或修改样本及其标签。图3-5展示了在交通标志上的典型投毒攻击形式。

图 3-5　投毒攻击

1. 投毒攻击分类

根据攻击者的攻击面、目标、知识、能力等不同角度可给出具体的投毒攻击分类[39]，表3-3对投毒攻击分类进行了总结。

表 3-3 投毒攻击分类

分类依据	分类结果
攻击面	数据投毒；模型投毒
攻击目标	有目标投毒；无目标投毒
攻击者知识	白盒攻击；黑盒攻击；灰盒攻击
攻击者的方法和能力	数据注入攻击；数据修改攻击；标签篡改攻击；模型篡改攻击

根据攻击者的攻击面（即训练阶段进行投毒攻击的目标），投毒攻击可以分为数据投毒和模型投毒。

- 数据投毒：在该攻击策略中，攻击者通过操控一小部分训练数据来影响模型的学习过程，尤其是模型的输入输出行为。这种情况常见于使用外包或未经充分验证的第三方数据集，包括互联网收集的数据。数据投毒攻击细分为两种类型：脏标签投毒和干净标签投毒。脏标签投毒涉及故意标错部分训练样本的标签，而干净标签投毒则保持标签正确，但修改样本内容以误导模型。
  - 脏标签数据投毒攻击：该攻击中，攻击者需要权限去修改数据标签和一部分训练数据的内容。
  - 干净标签数据投毒攻击：此攻击策略中，攻击者精心修改部分样本的内容而不改变其原有标签，然后将这些样本注入训练集中。其关键在于，尽管样本内容已被篡改，但其标签仍被正确保持，通常是由数据收集者或初步的数据处理者进行标注。这种方式的隐蔽性使得模型在训练过程中难以识别并排除这些恶意样本。
- 模型投毒攻击：该投毒攻击直接针对机器学习模型，它发生在攻击者能够控制或操纵学习算法及其流程的情况下。例如，攻击者拥有对模型训练过程的高级访问权限，而模型的使用者或开发者无法完全监控或控制这些流程，这使得攻击者能够植入恶意代码或修改模型的训练算法，从而导致模型输出预期之外的结果。

根据攻击者的目标，从对抗意图的角度来看，投毒攻击可以分为以下两类。

- 有目标投毒攻击：该攻击策略在受害者的模型投入使用时发生，旨在实现特定的系统逃逸目的。具体来说，攻击者精确了解了某个特定的测试数据集合 $X_t \in D_{\text{test}}$ 并预计这些数据将被用于模型测试。攻击者的目标是通过操纵这些数据集 X_t，使得模型在预测时产生错误或偏差的结果。
- 无目标投毒攻击：此类攻击中，攻击者的目标是实现一种拒绝服务攻击，通过尽可能增多目标模型的错误分类，来降低目标模型的可用性。

无目标投毒攻击在无人系统等实际应用场景中容易被检测并难以成功实施。因此，本节将主要聚焦于有目标投毒攻击的介绍和讨论。如无特殊说明，所提到的投毒攻击均为有目标投毒攻击。

知识在很大程度上限制了攻击者的能力和攻击策略。根据攻击者的知识，可以将投毒攻击分为以下三类。

- 白盒攻击：在该攻击场景中，攻击者完全了解目标机器学习模型 $M = (D_{\text{train}}, f, w)$，$D_{\text{train}}$ 是训练集，f 是学习算法，w 是模型内部参数。他们明确地知道目标模型的学习任务并且知道训练者选择了什么样的数据集和学习算法，还可以直接访问训练数据和内部模型参数。

- 黑盒攻击：在该攻击场景下，攻击者无法直接访问目标机器学习模型的内部参数或训练数据集。尽管如此，攻击者并非对受害者一无所知；他们了解模型的主要任务，并能够获得模型的输入和输出数据。
- 灰盒攻击：灰盒攻击是一种介于白盒和黑盒之间的复杂情况，攻击者对受害者有部分了解，即可能了解受害者的训练数据和模型类型。

在黑盒和灰盒场景中，由于攻击者有能力推断出受害者选择的算法类型和训练数据分布，基于这个假设，攻击者可以通过从互联网上收集训练数据，根据算法类型得到一个和目标模型接近的替代模型 M'，从而实现从黑盒攻击或灰盒攻击到白盒攻击的转变。

根据攻击者执行攻击的方法和攻击者的能力，投毒攻击可以分为以下四类。

- 数据注入攻击：攻击者能够通过将有毒数据注入训练集，来实现投毒攻击。
- 数据修改攻击：攻击者能够访问训练集，并且可以直接修改训练数据，从而实现投毒攻击。
- 标签篡改攻击：攻击者能够操纵数据标记过程，将错误的标签分配给训练样本，从而实现投毒攻击。
- 模型篡改攻击：攻击者能够控制模型训练算法和训练过程，他们甚至可以篡改模型的权重和系统中一些必要的文档，以实现投毒攻击。

前三类投毒攻击通常是数据投毒攻击，最后一类投毒攻击通常是模型投毒攻击。

2. 典型投毒攻击方法

- 标签投毒攻击：这种攻击是数据投毒策略中最常见的形式之一。在该攻击中，攻击者故意扭曲训练数据的标签，目的是建立错误的数据-标签关联，从而误导机器学习模型并降低其性能。特别是在标签翻转攻击中，例如二元分类问题，攻击者通过简单地翻转标签 0 和 1 来影响模型的学习过程。这通常需要攻击者能够访问和操纵标签，因此，这类攻击往往需要攻击者具备白盒知识。例如，攻击者可通过优化方法选择能最大化分类错误的数据点，显著降低支持向量机和逻辑回归模型的准确性[40]。攻击者可利用投影梯度上升算法针对黑盒学习模型进行标签投毒攻击，由于攻击者只需要黑盒知识，因此该攻击在实际场景中更为实用[41]。
- 基于优化的投毒攻击：基于优化的投毒攻击也是一种标签投毒攻击，可以帮助计算用于标签投毒的最佳数据集，同时可用于找到最有效的数据修改或数据注入方案。基于优化的投毒攻击的核心在于优化方程，将待解决的问题总结为一个最大化或最小化方程，该方法的通用工作流程如下。

  （1）攻击者首先将投毒问题转换为一个优化函数，以找到全局最优值。

  （2）攻击者使用梯度下降等优化算法在相应的约束下寻找解决方案。基于此双层优化问题的投毒攻击的性能主要取决于优化方程的构建及解决优化问题的策略，并且可以通过梯度方法高效解决双层优化问题，从而实现投毒攻击[42]。基于优化的投毒攻击框架一般通过改变目标函数、攻击目标和训练数据集的选择，几乎可以用于投毒攻击的所有场景和应用。

$$\begin{aligned} D_p^* &\in \arg\max_{D_p} \mathcal{F}(D_p, w^*) = \mathcal{L}_1(D_{\text{val}}, w^*) \\ &\text{s.t. } w^* \in \arg\min_{w} \mathcal{L}_2(D_{\text{train}} \cup D_p, w) \end{aligned} \tag{3.10}$$

其中，D_{train} 是原始训练集，D_{val} 是原始验证集，D_p 是一个中毒样本的集合。外层优化函数的目标是找到最有效的一个投毒样本集合 D_p^*，使得最大化投毒模型 w^* 在验证集 D_{val} 上的经验损失函数 $\mathcal{L}_1(D_{\text{val}}, w^*)$；内层优化则聚焦于使用原始训练集与中毒样本的并集来训练和调整模型参数 w^*。通过合适地选择损失函数 $\mathcal{L}_1$ 和 $\mathcal{L}_2$，该两层优化框架可以实现针对各种机器学习系统的投毒攻击。

- P 篡改攻击：P 篡改攻击允许攻击者在有限的预算 P 下向训练数据附加恶意的带毒噪声，这意味着任何传入的训练样本都可能以独立的概率 P 而受到投毒篡改。P 篡改攻击属于有目标的投毒攻击，尤其是在在线学习领域；在 P 篡改攻击中，攻击者具有数据修改和数据注入的能力，但是不能修改任何样本的标签。一种典型的 P 篡改投毒攻击可通过偏移任何有界实值函数的平均输出，从而增加训练模型在特定测试样本上的损失[43-44]；近期研究表明，多方学习过程中（如联邦学习）也可能受到 P 篡改投毒攻击的影响，增加模型在攻击者期望的特定目标样本上分类失败的概率[45]。
- 基于梯度的投毒攻击：基于梯度的方法通常通过迭代的步骤来找到可微函数的全局最优值。基于梯度的攻击通过朝着对抗目标函数的梯度扰动数据，直到投毒攻击产生最大影响，具体而言，可以通过基于梯度的优化方法解决式（3.10）中的双层优化问题。由于梯度的计算存储需要大量的资源，因此，针对梯度投毒攻击的研究侧重于改进计算密集型策略，以克服投毒中计算资源的限制。梯度反传方法作为一种更有效率和稳定性的方式来计算内部优化梯度，可成功对神经网络实施投毒攻击[46]。研究表明，通过快速海森向量乘积和共轭梯度求解器的组合，结合梯度反传方法，可有效解决投毒两层优化问题[47]。
- 基于生成模型的投毒攻击：传统的投毒攻击受限于中毒样本生成的速率，而利用生成模型绕过双层优化中的梯度计算，可以加速中毒样本的生成过程。同时，生成模型可以学习对抗性带毒扰动的概率分布，进而大规模制造中毒样本；生成模型需要对目标模型具有有限甚至完全的知识，分别对应于灰盒攻击和白盒攻击。例如，可通过构造一个编码-解码框架来加速投毒攻击的流程，该框架利用一个生成器不断产生带毒扰动，同时通过一个带毒分类器来检测模型的对抗性[48]。此外，可通过生成对抗网络来生成中毒样本，鉴别器的目标是区分带毒和良性样本，生成器的目标是生成中毒样本，其优化目标为最大化目标分类器的误差，最小化鉴别器将其生成的样本和良性样本区分的能力，通过对抗性生成中毒样本，实现在攻击强度和攻击隐蔽性之间的平衡[49]。
- 干净标签投毒攻击：干净标签投毒攻击不要求攻击者对于标记的过程有任何控制，它是一个更加现实的假设。干净标签投毒攻击通过生成微小的扰动来构造中毒样本，然后将其注入训练集之中，而不主动更改它们的标签。由于这些带毒图像看起来与未经修改的图像无异，人工审查员会根据它们的外观为每个样本贴上标签。在机器学习模型部署后，带毒的图像通常会影响分类器对特定目标测试样本的行为，而不影响其他测试样本上的行为，因此使得干净标签投毒攻击非常难以检测。下面介绍一种称为特征碰撞的毒化策略[50]：

$$x_p = \arg\min_{x_p} \|f(x_p) - f(x_t)\|_2^2 + \beta \|f(x_p) - f(x_b)\|_2^2 \tag{3.11}$$

其中，x_p 表示中毒样本，x_b 表示一个训练集中的基础样本，x_t 表示目标测试样本，f 表示受害者的特征提取器。式 (3.11) 的目的在于令中毒样本和目标样本的特征分布尽量一致，同时通过限制中毒样本与基础样本之间的距离，确保中毒样本在视觉上仍然正常，从而能够被正确标记。

- 基于影响函数的投毒攻击：通常地，选择投毒的样本不同可能导致不同的攻击性能，因此，引入影响函数来提高投毒攻击的强度。影响函数是鲁棒性统计学的一种经典技术，展示了当训练样本微小变化时模型参数的变化，基于影响函数的投毒攻击可以帮助理解中毒样本对模型预测的影响，从而用于构造最高效的投毒攻击[47,51]。
- 后门攻击：在训练阶段，攻击者通过数据投毒的方法，利用带有触发器（诸如特定的图案、标记）的训练数据将隐藏的后门嵌入神经网络之中，攻击者可以通过修改带有触发器的样本的标签，实现后门注入[52-53]。在测试阶段，面对良性样本，模型后门不会被激活，模型表现正常；而面对带有触发器的中毒样本，模型中的后门将被激活，会将该样本分类到攻击者提前指定的类别之中，因此，后门攻击是一种隐蔽性很强的有目标数据投毒攻击。
- 本地模型投毒攻击：分布式学习可能会受到恶意参与者的投毒攻击，攻击者伪装成良性参与者，但上传毒化更新给中心聚合服务器，从而轻易影响全局模型的性能[54-55]。分布式学习框架为敌手进行模型投毒攻击提供了很多可操作的漏洞，一方面，在这个框架中，本地模型的训练样本不会被释放给可信的权威机构进行检查，因此，服务器无法确认客户端的更新是否真实和正确。另一方面，在该设置下，服务器很难区分来自对手的恶意更新和来自正常客户端的良性更新，因此难以检测分布式学习中的模型投毒攻击。

3.4.2　投毒攻击防御方法

投毒攻击揭示了数据驱动的机器学习模型训练过程中数据供应链的脆弱性，它强调了设计能抵御投毒攻击的机器学习系统以提高模型的鲁棒性的重要性。表3-4总结了投毒攻击的主要防御方法。

表 3-4　投毒攻击防御方法及其主要技术

防 御 方 法	主 要 技 术
中毒样本过滤（被动）	离群值过滤；基于集成的过滤；基于验证的过滤；基于 k 近邻的过滤
模型净化（被动）	客户端交叉验证；谱异常检测；验证数据集
数据预处理（主动）	数据增强
鲁棒性训练（主动）	鲁棒性过程；鲁棒性结构

根据防御者是否在事后采取补救行动或是在事前采取预防行动，可以将投毒攻击的防御方法分为被动防御和主动防御，其中被动防御又可以分为中毒样本过滤和模型净化的方法，而主动防御可以分为数据预处理和鲁棒性训练。

- 被动防御
  - 中毒样本过滤：防御者通过过滤算法清理训练数据集，以识别、移除或修正中毒样本。

  - 模型净化：防御者通过对受害模型内部参数检测异常，并修正其异常行为使其正常运行。
- 主动防御
  - 数据预处理：防御者通过数据清洗、增强和转换来对原始训练数据进行预处理，以防止投毒攻击。
  - 鲁棒性训练：防御者通过增强训练阶段模型的鲁棒性，使得投毒攻击难以成功。

被动防御侧重于检查已经实现的攻击，并对损失进行弥补。针对数据投毒攻击，只要在训练之前对中毒样本进行过滤，训练的模型就可以保证是没有被污染的；然而，模型投毒攻击将直接污染模型的参数，防御者必须进行模型净化，这增加了被动防御的挑战性。

1. 被动防御：中毒样本过滤

大多数数据集中不可避免地存在“脏数据”，这些“脏数据”可能包含异常事件或者恶意攻击引起的噪声，从而模糊了这个样本的特征和其对应标签之间的关系。当数据被投毒攻击污染时，通过过滤方法改善训练数据的质量是一种直观的解决方案，类似于离群值检测或者异常检测，实现针对中毒样本的识别，进而在训练过程开始之前实现重新标记或者移除。

- 离群值过滤：离群值过滤是最简单的投毒攻击防御方法之一，在一般情况下，中毒样本表现出的行为和离群值一样，通常远离每个类的质心，因此，消除这些离群异常值以防止数据中毒是一种直观的想法。通过计算每个类的质心，并且去除远离相应类质心的点以实现离群值过滤，从而可有效防御数据投毒攻击[56-57]。然而，如果中毒样本占比较多，每个类的质心可能都会偏离正确的路径，此时进行离群值过滤反而可能对模型性能造成更大的损害。
- 基于集成的过滤：集成方法的目标是将来自多个假设的预测集成以形成一个更好的预测，具体而言，防御者通过在原始数据集的不相交子集上进行训练以生成多个基础模型，在这些基础模型上训练测试训练集中的每个样本。如果大多数基础模型的预测与观察到的标签不同，那么认为该样本很可能是一个中毒样本[58]。然而，如果基础模型不能准确区分良性和中毒样本，数据不平衡可能导致过多良性样本被错误移除，从而降低模型性能。
- 基于验证的过滤：基于验证的过滤依赖于干净且标记正确的验证数据集，可通过验证数据集训练区分分类器或辅助模型来有效缓解投毒攻击[56,59]。然而，如果验证数据量有限，该方法表现可能难以达到预期。
- 基于k近邻的过滤：使用独立训练的k近邻分类器来检测中毒样本，将每个训练样本的标签与特征分布中的k个最近邻继续进行比较，随后消除或重新标记那些与其邻居标签不一致的可疑样本[56,60]。

2. 被动防御：模型净化

本节以分布式机器学习框架联邦学习为例，来说明被动防御中的模型净化技术。因为本地训练数据不对外公开，无法被服务器或任何第三方机构审查，基于中毒样本过滤的防御方法在联邦学习中不适用，因此，防御者必须检查所有来自本地的更新，以保护全局模型不受攻击。客户端交叉验证技术可评估各个更新在其他客户端的本地数据集上的表现，根据这些评估结果，服务器在聚合时动态调整每个更新的权重[61]。基于谱异常检测的框架可根据其低维嵌入区分异常的模型更新和良性的更新，在谱异常检测之后，将删除异常更新，

只保留良性更新[62]。此外，研究者提出了一种基于验证数据集的模型净化方法，利用服务器上的验证数据集来训练模型，并根据每次更新与服务器模型的余弦相似度动态调整聚合权重[63]。

联邦学习中防御投毒攻击的主要思想是检查依赖于中央服务器的模型参数。基于分布式设置中的多源权重更新，来自良性用户的更新可以用于帮助识别恶意更新，从而提高分布式学习系统的鲁棒性和抗性。

对于未知投毒攻击的主动防御，一种有效的策略是采取积极的预防措施，这涉及在数据预处理阶段和模型训练方法上的改进，以增强系统对投毒攻击的抵御能力。

3. 主动防御：数据预处理

在数据预处理过程中，可以通过数据清洗、增强和转换对原始训练数据进行预处理，从而使得模型更加鲁棒。针对投毒攻击的防御方法可能会因剔除大量良性训练数据而导致模型性能显著下降，引起训练不平衡和过拟合。通过添加数据而非减少数据的方式同样可以使得模型的鲁棒性显著增加。数据增强方法可以在不牺牲性能的情况下，显著减弱投毒攻击的威胁[64]。前者能够对随机抽样的训练数据进行成对凸组合，并利用相应的标签凸组合进行标记，这种方法能够防止对带毒标签的记忆，从而提供了针对投毒攻击的鲁棒性，同时提高了模型的泛化能力。后者通过在一幅图像中随机选择补丁，并将这些补丁叠加到其他图像上。接着，标签根据补丁面积的比例进行混合，从而增强模型的鲁棒性并提高测试准确性。总体而言，这些数据增强策略有效地防御了数据投毒攻击，同时保持了高水平的自然验证准确性。

4. 主动防御：鲁棒性训练

鲁棒性指的是系统在执行过程中处理异常和错误输入的能力。因此，鲁棒训练的目标是使训练阶段更加健壮，从而使毒害攻击更难成功。

- 鲁棒性过程：对于结构比较简单的传统机器学习系统，防御者通过部分过程修改，如调整损失计算、特殊处理参数等提高鲁棒性[65-66]。
- 鲁棒性结构：深度神经网络在本质上比机器学习更加鲁棒，但仍然相当容易受到毒化攻击的影响。在这种情况下，防御者通过向原始网络添加结构模块来提高模型的鲁棒性[67-68]。

3.4.3　无人系统中的投毒攻击

无人系统容易受到投毒攻击的影响，投毒攻击可以误导无人系统分类模型的结果，如交通标志识别，从而影响自动驾驶系统的决策，进而导致不可预测的危险[2,69]。以交通标志识别任务为例，自动驾驶汽车利用深度神经网络对摄像机捕捉的交通标志图像进行分类，并利用智能控制系统根据预测结果控制智能汽车，该过程中，攻击者可以通过多种方式将中毒样本注入自动驾驶汽车模型。首先，这些深度学习模型使用大量从汽车用户收集的训练样本，难以确保所有样本均为良性。攻击者可以在收集阶段向原始训练集上传中毒样本，以影响模型结果。其次，这些模型需要根据新收到的训练样本定期进行更新，以应对不断生成的数据。当模型训练者使用带毒样本重新训练模型时，即使经过良性的训练过程，模型仍可能被感染。在推断时，中毒模型将目标样本误分类为指定的分类，例如，在样本肉眼看上去交通标志为“停止”时，但却决策为“加速”。

面向自动驾驶的一个具体投毒攻击案例如下：通过构建并放置正方形或者苹果公司标志之类的触发器，在原始输入图像的角落实施针对端到端自动驾驶的后门攻击，如果道路图像包含这些恶意触发器，车辆很可能会偏离预先规划的轨迹[2]，同时仅通过观察测试准确度的结果很难判定模型是否遭受了模型投毒攻击[70]。此外，可利用生成模型进行中毒攻击，诸如利用如GAN从车前玻璃图像中去除雨滴，当中毒样本被注入原始训练数据中，使得生成模型学习到从输入域到输出域的错误映射，当生成模型去除雨滴时，同时将红色交通灯变为绿色，或者改变限速标志上的数字，均可能对自动驾驶系统造成误导[71]。针对激光雷达感知系统，存在针对激光点云的后门攻击，具体地，对于某些激光雷达感知系统开发商或者自动驾驶汽车，从不同的来源（诸如不同的自动驾驶汽车用户）收集训练数据或者将训练工作外包给第三方，均导致攻击者有机会通过针对目标模型进行投毒攻击，将隐藏的触发模式注入受害者检测模型中；同时攻击者可以对少量点云样本进行投毒攻击，具体地，通过创建假车辆点簇并将它们隐藏在点云中，从而实现隐蔽性强的后门攻击[72]。针对无人机系统存在一种新的数据投毒攻击，具体地，攻击者针对无人机内部架构，通过有毒碰撞和转向角数据集，破坏无人机的自主导航系统模型，降低自主导航系统中神经网络模型的准确率，同时通过转向角分布图可视化和碰撞数据集饼图可视化的检测技术，可以防御数据投毒攻击[69]。自动驾驶汽车和无人机等无人系统中还存在一种地图数据投毒攻击，攻击者能够通过发送带毒信息以影响无人系统的地图数据库的精度，从而造成误导[73]。

3.5 多源数据融合安全

无人系统数据融合技术作为一种信息智能处理方法，集合了来自各类传感器、硬件设施及不同信息平台的多元数据资源，旨在全面理解系统所处环境的状态、任务需求及运行状况。通过深度融合各类数据，无人系统能够在更高的精度水平上认知周边环境，制定出更为精确和优化的决策方案，从而全面提升系统的整体效能和工作效率，达到超越单一传感器或信息来源所能提供的推理和判断能力。例如，在自动驾驶无人车辆场景中，数据融合技术扮演着至关重要的角色。车辆采集来自摄像头、激光雷达、GPS定位系统等多种传感器的数据信息，通过深度整合和协同处理多模态数据，自动驾驶车辆能够更准确地识别道路环境，从而实现更安全稳定的驾驶。相较于仅依赖单一传感器的数据分析，多源数据融合后的结论往往具备更高的准确性和可靠性，为无人系统的自主路径规划、决策制定和运动控制提供坚实的数据支撑。图3-6展示了数据融合的基本流程。

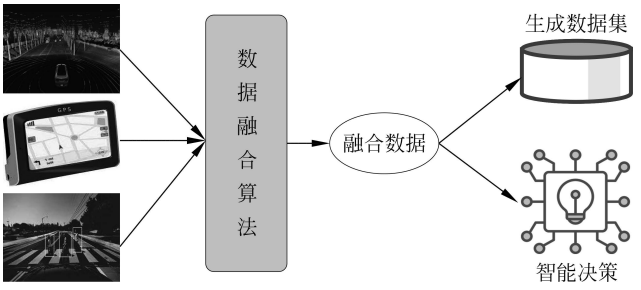

图 3-6　数据融合示意图

3.5.1　数据融合方法

典型的数据融合方法大致可以分为三类：基于概率的方法、基于证据推理的方法和基于机器学习的方法。

（1）基于概率的数据融合方法。基于概率的数据融合方法通过引入概率分布函数或者概率密度函数来处理数据的不完美，通过随机变量之间的依赖关系并建立不同数据集之间的关系，从大量的冗余数据中提取所需的关键特征。常见的方法如下。

- 贝叶斯推断：贝叶斯推断利用贝叶斯公式整合多个信息源的不确定信息。它通过更新后验概率，从先验概率和条件概率中提高对目标的可信度判断。尽管贝叶斯推断能高效地融合先验知识，但它在处理高维数据时的计算复杂度较高，且对先验概率的选择较敏感。
- 卡尔曼滤波：卡尔曼滤波结合了不同传感器的预测与观测结果，实现了精确的后验估计。这种方法能高效地实现实时数据融合与预测。然而，它不适用于非线性系统，且在处理高维状态空间时面临计算复杂性挑战。
- 马尔可夫模型：马尔可夫模型可以通过结合不同传感器的信息，更新状态转移概率和观测模型，从而提供准确的状态估计，其优点在于可以简单地实现序列数据处理，缺点在于对于初始条件设定敏感，同时在高维状态空间中存在计算复杂性高的问题。

（2）基于证据推理的数据融合方法。基于证据推理的数据融合方法的主要理论基础是Dempster-Shafer (D-S) 证据理论。该理论通过引入主观信念函数来处理不确定信息，从而有效地表达现实世界的不确定性并使得证据推理在动态场景中得以应用，它的优点是能在信息缺乏的情况下高效地处理不确定性，主要缺点是在非线性场景下难以确定质量函数。

（3）基于机器学习的数据融合方法。机器学习可以基于已知数据进行分类和预测，并能够挖掘数据输入中的深层关系。近年来，机器学习已被广泛应用于数据融合领域，显著提高了数据融合方法的性能，常见的基于机器学习的数据融合方法可大致分为有监督机器学习数据融合、无监督机器学习数据融合以及模糊逻辑数据融合，具体说明如下。

- 有监督机器学习：通过k近邻、支持向量机（Support Vector Machine，SVM）、神经网络等有监督机器学习算法，将来自多传感器捕获的原始数据处理后作为算法输入，通过机器学习算法进行数据融合，输出具有高精度、高可靠性的数据。
- 无监督机器学习：无监督机器学习使用如k中心、k均值、DBSCAN等聚类算法，将来自不同来源的数据分组成簇。这些方法通过将相似的数据点集中处理，实现数据的有效合成，进而输出融合后的结果。
- 模糊逻辑：多传感器数据融合中，模糊逻辑对传感器输出进行模糊化处理，将测量值按不同级别分级并表示为相应的模糊子集。通过设定隶属函数并使用多值逻辑，这些函数被综合处理以输出一个明确的融合值。

3.5.2　多源数据融合需求

在无人系统中，多种传感器被部署用于感知或收集数据，通过进一步的数据融合和分析来提供对环境的理解，从而便于无人系统进行高效智能决策。无人系统中的感知数据具有如下特征。

- 数据量大：无人系统中，多类传感器（诸如摄像头、雷达等）对于周围环境不断进行感知和检测，产生了大量数据。由于传感器设备的存储和计算能力有限，无人系统中的数据融合面临难以区分虚假数据、多源数据导致的通信负担增加、融合准确率受到对不同传感器数据的信任过度或不足的影响等挑战。
- 多模态：无人系统包含多种传感器，产生的多模态数据因其类型、形式、表示、尺度和密度的差异，使得数据融合极为复杂。
- 隐私敏感：无人系统中传感器感知到的信息可能包含无人系统用户的隐私信息，例如，在自动驾驶或者无人机系统中，感知到的数据可能包括用户的位置和偏好等，存在隐私泄露的风险，多传感器数据融合更加剧了这种风险，在数据融合中保护隐私的同时确保融合准确性是具有挑战性的。
- 可靠性低：无人系统中的各种传感器可能遭受环境影响及受到恶意攻击，因此传感器感知到的数据可能是不精确、不确定的，数据的低可靠性为多源数据融合带来了额外的挑战。
- 动态变化：无人系统中的多种传感器随着环境变化实时进行感知，因此感知数据的动态变化和新鲜度也是影响多源数据融合质量的关键因素。

根据无人系统中感知数据的特征，可以总结出无人系统中安全可靠的多源数据融合需求如下。

上下文感知：多源数据融合应具备上下文感知能力，以支持高度智能的环境自适应和服务灵活性。其中，上下文指用于描述涉及实体的背景或情境的信息，而具备识别和适应上下文的能力即为上下文感知。上下文信息可能不是固定的与实体直接关联，并可能随时间变化。以位置信息为例，作为最常用的上下文信息之一，它决定了在自动驾驶路径规划中应该考虑哪些传感器。

隐私保护：在多类传感器设备收集和传输环境数据进行进一步数据融合与分析的过程中，可能会涉及敏感且隐私的信息。例如，在无人机系统中，传感器可能会记录用户的位置或偏好等私人信息，如果不采取适当的隐私保护措施，这些信息可能因隐私窃听而泄露。

可靠性：多源数据融合的结果通常直接导致特定的决策，如紧急响应、路径规划等。不可靠的融合结果可能令无人系统用户处于危险之中，可靠性是多源数据融合的最基本、最重要的要求之一，在融合结果进入决策系统之前，需要对融合结果进行可靠性评估。

实时性：无人系统的环境随着时间快速变化，需要短时间内进行实时数据融合和分析，以实现无人系统对于环境变化，特别是紧急情况下的快速响应。高效的数据融合能够优化融合中心或传感器的计算通信开销，进一步支持对于融合结果分析的快速响应，从而提供更好的用户体验。

鲁棒性：在无人系统中，数据融合可能遭受敌手的多种数据攻击（诸如虚假数据注入攻击），因此数据融合应具有抵抗各种攻击的鲁棒性，否则，各类攻击将加剧收集到的各类数据的不确定性，进而产生不正确的融合结果。

可验证性：无人系统或用户应可以验证数据融合的结果，一方面，这个要求使得用户能够评估数据融合结果的正确性；另一方面，可验证性可以帮助无人系统检查所收集数据的质量。例如，具有可验证性的数据融合可以帮助判断一条感知到的环境数据是否实际对最终决策产生了影响。

3.5.3　多源数据融合关键技术

为了实现安全可靠的多源数据融合，首先需要通过数据清洗、融合后的质量评估和验证确保融合结果的正确性和算法的鲁棒性；其次，保护过程中和结果中的隐私性至关重要，这需要在数据传输和存储中应用加密技术，并通过可信计算进行隐私数据融合，以及实施针对融合结果的细粒度访问控制。图3-7展示了多源数据融合流程及相应关键技术在数据融合的不同生命周期中的应用。

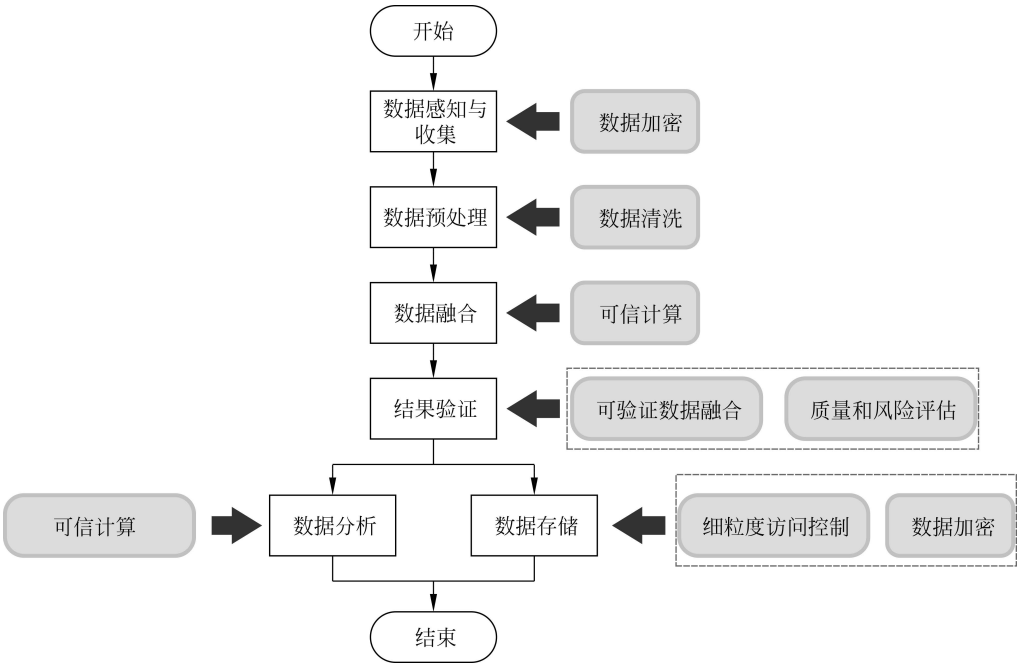

图 3-7　多源数据融合流程及其关键技术

（1）数据加密。密码学方法可以保证传感器数据的机密性和安全性能够有效地确保数据融合技术的隐私保护能力。对于传感器感知到的用于数据融合的数据，在数据传输过程中进行加密处理，通过安全协议进行传输；对于数据融合之后的结果，在数据存储时采用加密算法进行处理，防止无人系统被攻击导致隐私信息泄露。具体而言，传感器应采用SSL/TLS协议用于数据安全传输，并配合AES、RSA等加密算法，保障数据在传输过程中的安全。

（2）数据清洗。数据清洗一方面可以通过不良数据检测和数据降噪，提高感知数据的可靠性，进而提高数据融合结果的准确性和可靠性；另一方面可以对敏感数据进行数据脱敏，对敏感信息进行不可逆的处理，从而实现敏感数据的隐私保护。数据脱敏作为确保系统安全性和用户隐私的关键步骤，能够通过对敏感信息进行替换、模糊、遮挡等方式进行处理，从而避免在数据融合中泄露敏感数据。例如，在自动驾驶系统中，可以对摄像头捕获的敏感区域图像进行模糊处理或添加阻挡物，尤其是在涉及行人或车辆等隐私敏感区域。其中，数据脱敏的程度应在不影响无人系统性能的前提下，保证数据隐私。

（3）可信计算。可信计算技术指在计算机系统中保障计算过程和计算结果的可信性、完整性和机密性的一种技术。具体而言，在数据融合中，可信计算通过硬件安全技术、软件安全技术及安全协议等，在保证原始感知数据隐私的情况下，完成数据融合计算。同态加密是

实现可信计算的重要技术之一。在数据融合领域，通常利用同态加密技术实现隐私保护的数据融合。尽管同态加密可以用于隐私保护的数据融合，但高昂的计算成本使其难以有效应用于实时无人系统场景中；在高度分布式地系统中，基于秘密共享的方案成为同态加密数据融合方案的一种可能的替代方案，其能够将数据在无人系统中进行融合，并降低了通信成本。

（4）细粒度访问控制。数据融合结果通常比原始数据包含更多有意义的信息，因此细粒度访问控制成为无人系统的基本要求，考虑到数据融合隐私和安全需求，有必要对多源数据融合后的结果进行安全灵活的细粒度访问控制。细粒度访问控制可以通过属性基加密的方式实现，可以通过向特定的授权用户组共享融合结果来保护用户的隐私，降低隐私泄露的风险，诸如文献[74]就提出了通过属性基加密实现可撤销的细粒度访问控制。

（5）可验证数据融合。数据融合的可验证性有助于追溯问题的根源，并定位针对数据融合的攻击。为了保证数据融合结果的正确性，需要在确保原始数据的隐私的情况下实现数据融合。同态签名是验证数据融合结果的一种有效技术，能够在数据融合前后保证数字签名的有效性，能够在不暴露原始数据的情况下，进行有效的融合验证。区块链技术是实现可验证性数据融合的另一个关键技术，可以确保整个数据融合过程的透明性和可追溯性。区块链中可进一步融合隐私保护算法（如同态加密、差分隐私等）和智能合约，实现在不泄露原始数据的情况下进行数据融合，从而确保数据融合结果可验证性的同时保障敏感数据的机密性。

（6）质量和风险评估。在多源数据融合系统中，质量与风险评估是确保系统有效运作的重要方面。质量评估涉及对融合数据的准确性、完整性、一致性和时效性等方面进行评估，例如，验证融合结果是否与实际情况一致，数据是否包含错误或缺失，以及融合过程是否能够按时产生可信赖的输出，其目的是确保系统生成的融合数据是高质量的、可靠的且符合预期的。风险评估关注系统实施中可能出现的潜在问题和威胁，包括安全漏洞、隐私泄露、数据失真、技术可行性等方面的风险，风险评估的目的是在实施前识别、量化和管理潜在的问题，以制定有效的风险缓解措施，确保系统在运行时不会面临不可接受的威胁和风险。质量评估能够确保融合的数据是可信赖和高质量的，提高决策和分析的准确性；风险评估有助于降低在系统实施和运行过程中可能发生的各种问题，确保融合数据的安全性和隐私性，确保数据融合系统的稳定性。两者共同保证了多源数据融合系统的可靠性和安全性。

3.6 案例分析

3.6.1 背景简介

联邦学习是一种新兴的协同机器学习范式，旨在解决“数据孤岛”问题，同时保护参与者本地数据的隐私。联邦学习可作为无人系统中实现可信分布式计算和隐私保护模型训练的重要技术，能够让自动驾驶车辆、无人机等无人系统共同训练全局模型，而无须与中央服务器共享其本地数据。在通常的联邦学习过程中，参与者使用中央服务器广播的全局模型在其本地私有数据集上进行训练，生成相应的本地更新，并将其上传到服务器；而中央服务器使用平均聚合策略，即对接收到的上传数据取平均值，生成用于更新全局模型的聚合模型。

然而，联邦学习中可能存在拜占庭参与者，可能通过上传恶意更新来实施模型投毒攻

击，从而阻止全局模型正常收敛并且降低模型性能。研究表明，对于经典的模型投毒攻击，经典的平均聚合具有脆弱性[75]，即使只有一个拜占庭参与者，全局模型的性能也会受到影响。除了模型投毒攻击，模型训练过程中还存在的隐私威胁，诸如“诚实”但“好奇”的中央服务器。服务器可能会“诚实”地执行联邦学习步骤，聚合参与者的隐私更新，但对参与者的隐私数据产生“好奇心”，并通过梯度推断攻击试图推断敏感信息。研究表明，通过梯度推断攻击可以从更新或梯度中提取原始训练数据，造成严重的隐私威胁。

目前，学术界已经提出了几种针对模型投毒攻击的鲁棒联邦学习方案，主要分为两类：基于统计的方案和基于性能的方案。基于统计的方案通过排除与良性更新几何距离较远的更新来维护联邦学习的拜占庭鲁棒性，例如 Krum[75]、Median[76]、Trimmed mean[76]等。基于性能的方案通过减小性能较差的拜占庭参与者的更新对模型的负面影响。强化学习[77] 和余弦相似度比较[63] 是实现基于性能的拜占庭鲁棒联邦学习方案的两种主要技术。然而，所有现有的拜占庭鲁棒联邦学习方案都假设中央服务器是完全值得信赖的。它可以访问参与者的原始本地模型更新，并执行隐私攻击以提取参与者的敏感信息。此外，现有的方案在参与者数据高度非独立同分布（non-IID）的情况下会出现明显的性能下降。

此外，为了保护本地模型更新免受“诚实”但“好奇”的服务器攻击，在联邦学习过程中参与者会对其原始更新进行掩码或加密。然而，因为服务器无法在没有原始本地模型更新的情况下计算统计信息或评估更新的性能，本地模型更新的掩码或加密使得检测拜占庭攻击者变得困难。为了同时实现对本地参与者模型更新的隐私保护及针对模型投毒攻击的鲁棒性，需要实现在隐私情况下针对参与者模型更新的质量评估。当前方案主要使用安全多方计算（Secure Multi-Party Computation，MPC）和可信执行环境（Trusted Execution Environment，TEE）两种方式来实现隐私质量评估。MPC 允许所有参与者计算加密或掩码更新之间的欧几里得距离，用以检测被投毒的更新。此外，MPC 还支持用于聚合加密更新的拜占庭鲁棒方案。然而，在数据分布高度非独立同分布的情况下，现有方案无法保证针对多种攻击的有效防护。

本案例提出了 DRL-PBFL[78]，是一种通过深度强化学习（Deep Reinforcement Learning，DRL）来实现隐私保护的拜占庭鲁棒联邦学习方案。DRL-PBFL 能够有效抵御各种模型投毒攻击，包括针对该防御方案的定制投毒攻击，同时保证参与者的本地数据隐私，即防止“诚实”但“好奇”的服务器的梯度推断攻击。其主要思想是使用基于性能的加权聚合策略来聚合参与者的扰动后更新。本案例中，中心服务器端有一个具有有限大小的验证数据集，该数据集符合标准的数据分布，并用于评估更新的性能。直观地说，基于性能的加权聚合策略可以在联邦学习中高效识别拜占庭参与者并对其分配低权重，使得最终的全局模型能有效防御来自拜占庭参与者的模型投毒攻击。即使在参与者的数据分布高度非独立同分布的情况下，DRL-PBFL 方案也能有效地抵抗模型投毒攻击。聚合策略仅由全局模型在验证数据集上的性能决定，独立于参与者之间的本地数据分布，由于在验证数据集上无法直接评估掩码更新后的质量，利用深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）强化学习技术，根据在验证数据集上的前后全局模型表现，计算中心服务器的加权聚合策略。为了解决由权重引起的扰动消除问题，本案例设计了一种基于拉格朗日插值的安全聚合算法，参与者使用拉格朗日多项式和当前聚合策略构建扰动，这些扰动在服务器聚合更新后被消除。

3.6.2 威胁模型和防御目标

在本案例中，联邦学习的主要安全威胁来自“诚实”但“好奇”的中心服务器和拜占庭恶意参与者。

- “诚实”但“好奇”的中心服务器：虽然中心服务器在联邦学习过程中会“诚实”地执行每一步训练，但它也可能对参与者上传的更新抱有“好奇心”，并尝试提取其隐私敏感信息。通过梯度推断攻击，服务器有可能利用参与者在联邦学习中上传的梯度数据，从而提取私人原始训练数据。
- 拜占庭恶意参与者：对手可能会妥协所有参与者中的一部分，即ε的比例（其中$\varepsilon \in (0,1)$）。拜占庭参与者旨在执行无目标模型投毒攻击，发布恶意更新以降低全局模型的性能，导致在测试数据集上出现较高的错误率。拜占庭参与者可以任意操纵本地更新以执行不同的模型投毒攻击。

在本案例中，假设拜占庭参与者遵守服务器制定的训练协议。从这一协议偏离将破坏训练过程，并使拜占庭参与者易于检测。考虑到可能存在的服务器与参与者之间的共谋威胁，本案例假设至少有一个被其他参与者视为值得信赖的参与者，该参与者在扰动计算期间充当中继计算节点（如社交网络中的领导者），其应具有较高的声誉，并与服务器保持独立，不参与任何共谋。

DRL-PBFL的防御目标是防止中心服务器通过梯度推断攻击侵犯隐私，并防止拜占庭恶意参与者进行模型投毒攻击导致的性能下降，可以将防御目标形式化为

$$\begin{aligned} &\min \sum_{x\in\mathcal{D}_r} \mathcal{L}(\hat{\theta}_g(x)) \\ &\text{s.t. } \hat{\theta}_g = \theta_g + \mathcal{A}(f(\Delta\hat{\theta}_1),\cdots,f(\Delta\hat{\theta}_c),f(\Delta\theta_{c+1}),\cdots,f(\Delta\theta_n)) \end{aligned} \tag{3.12}$$

其中，θ_g是每个通信轮次中的全局模型，$\mathcal{D}_r$是服务器的验证数据集，$\mathcal{L}$是损失函数，$\mathcal{A}$是拜占庭鲁棒的聚合机制，而f是参与者应用于扰动本地更新的掩蔽方法。假定前c个更新来自拜占庭参与者。通过这种机制，服务器可以获得针对特定任务的有效模型而无须访问参与者的原始更新，从而同时保障服务器的效益和参与者的隐私安全。

3.6.3 方案总体设计

为了实现3.6.2节的防御目标，本案例采用了基于拉格朗日的加权安全聚合方案和深度学习技术DDPG机制，以实现隐私保护和拜占庭鲁棒性的联邦学习，称之为DRL-PBFL。图3-8展示了DRL-PBFL在每轮通信中的工作流程。DRL-PBFL遵循常见的联邦学习框架，并集成了基于拉格朗日的加权安全聚合方案和一个DDPG组件作为优化器来更新聚合策略。

聚合策略定义为与参与者对应的本地模型更新的权重，反映了模型质量和数据质量。全局模型在验证数据集$\mathcal{D}_r$上的性能决定了聚合策略的更新方式。随着训练的进行，权重将被动态调整，良性参与者将获得更高的权重，而拜占庭参与者的权重将被相应降低。基于全局模型在$\mathcal{D}_r$上的性能，DRL-PBFL通过DDPG强化学习机制给出新的聚合策略和优化后的参与者的权重，以便所有无目标模型投毒攻击都无法成功。然而，由于本地模型更新可能已经被本地参与者扰动，如果不考虑权重的影响，则无法在聚合后消除扰动。由于权重会改变

聚合步骤中原始扰动的大小，仅采用简单的加权聚合方法无法有效消除扰动。这使得传统的安全聚合方案在此场景下不再适用。为满足扰动消除的需求，DRL-PBFL 提出了一种基于拉格朗日插值的安全聚合算法。该算法根据当前聚合策略 $\mathcal{P}=\{w_1,\cdots,w_i,\cdots,w_n\}$ 生成可消除的扰动。

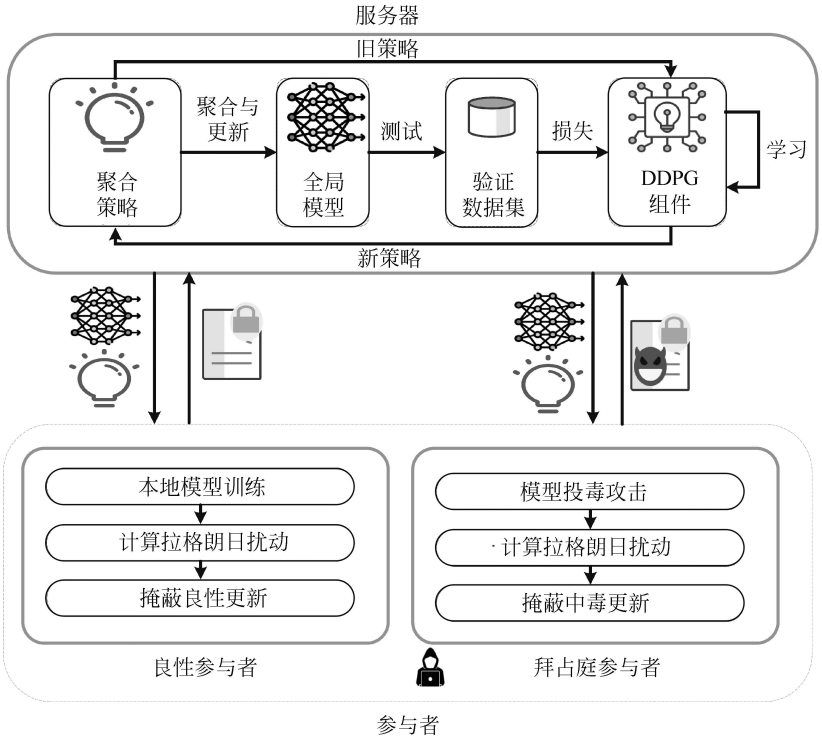

图 3-8　DRL-PBFL 工作流程

服务器首先初始化聚合策略 $\mathcal{P}$、全局模型 θ_g、验证数据集 $\mathcal{D}_r$ 和 DDPG 组件。在训练过程中，服务器与 n 个参与者之间进行多次通信交互。具体而言，每个训练轮主要包括以下步骤。

- 步骤 1：服务器广播全局模型 θ_g 和聚合策略 $\mathcal{P}$ 给参与者。
- 步骤 2：对于良性参与者，他们使用本地数据集训练本地模型并上传扰动后的更新；对于拜占庭参与者，他们通过生成各种模型投毒攻击并上传扰动后的拜占庭更新。这些扰动是根据当前聚合策略和拉格朗日插值方法计算得到的。
- 步骤 3：服务器接收所有扰动后的更新并根据聚合策略 $\mathcal{P}$ 进行聚合。接着，使用这些聚合后的更新来更新全局模型 θ_g。
- 步骤 4：服务器在验证数据集 $\mathcal{D}_r$ 上评估新的全局模型 θ_g'，并计算得到相应的损失 l_r。然后，将损失 l_r、旧策略 $\mathcal{P}$ 和辅助信息 $\mathcal{H}$ 输入 DDPG 组件，以生成新的聚合策略 $\mathcal{P}'$。
- 步骤 5：DDPG 组件通过内部强化学习过程更新其内部神经网络 $\{\mu,\mu',Q,Q'\}$。

对于本地训练，良性参与者首先创建一个本地模型 θ_l，该模型与原始全局模型 θ_g 相同，并使用随机梯度下降（Stochastic Gradient Descent，SGD）算法优化 θ_l 如下：

$$\theta_l'=\theta_l-\alpha\nabla_{\theta_l}J(\theta_l;x(i),y(i)) \tag{3.13}$$

其中，J是损失函数，α是学习率。然后，良性更新$\Delta\theta_b$计算如下：

$$\Delta\theta_b = \theta_l' - \theta_g \tag{3.14}$$

与良性参与者不同，拜占庭参与者使用模型投毒攻击生成恶意的本地更新。

然后，每个参与者根据拉格朗日插值通过聚合策略生成扰动p_i。每个参与者通过添加扰动p_i掩蔽更新$\Delta\theta_i$并上传掩蔽后的更新$\Delta\theta_i'$：

$$\Delta\theta_i' = \Delta\theta_i + p_i \tag{3.15}$$

服务器使用线性加权方法和当前轮次的聚合策略$\mathcal{P}$计算聚合更新$\Delta\theta$：

$$\Delta\theta = \sum_{i=1}^{n} w_i \Delta\theta_i' \tag{3.16}$$

$\Delta\theta$用于使用全局学习率α_g更新全局模型θ_g：

$$\theta_g' = \theta_g - \alpha_g \Delta\theta \tag{3.17}$$

更新后的全局模型θ_g'在$\mathcal{D}_r$上的损失l_r可通过下面的公式进行计算：

$$l_r = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \theta_g'(x_i)), \quad (x_i, y_i) \in \mathcal{D}_r \tag{3.18}$$

损失l_r、旧策略$\mathcal{P}$和辅助信息$\mathcal{H}$组合成为状态S，作为DDPG 组件的输入。辅助信息$\mathcal{H}$用于协助 DDPG组件在每个强化学习轮次的开始生成合理的聚合策略，$\mathcal{H}$可以是参与者的历史得分、服务器的主观评估等。然后，DDPG 组件输出新的聚合策略$\mathcal{P}'$：

$$\begin{aligned} \mathcal{P}' &= \mathrm{DDPG}(S) \\ S &= (l_r, \mathcal{P}, \mathcal{H}) \end{aligned} \tag{3.19}$$

随着训练过程的进行，DDPG组件更新聚合策略$\mathcal{P}$，降低拜占庭参与者在聚合步骤中的影响，并增强良性参与者的影响。

3.6.4 基于拉格朗日插值的安全聚合机制

本案例提出了一种基于拉格朗日插值的安全聚合算法，该算法根据当前的聚合策略$\mathcal{P} = \{w_i\}$生成可消除的扰动。具体而言，假设有n个参与者，每个参与者都收到聚合策略$\mathcal{P}$。这些参与者首先就一个常数项为0的$(n-1)$次多项式$f(x)$达成一致：

$$f(x) = c_1 x^{n-1} + c_2 x^{n-2} + \cdots + c_{n-2}x^2 + c_{n-1}x \tag{3.20}$$

其中，$c_1, \cdots, c_{n-1}$是$f(x)$的参数。然后，每个参与者i选择一个满足以下条件的秘密参数s_i：

$$w_i = c_1 s_i^{n-1} + c_2 s_i^{n-2} + \cdots + c_{n-2}s_i^2 + c_{n-1}s_i \tag{3.21}$$

根据拉格朗日插值法，拉格朗日基函数$p_i(x)$可以写成：

$$\begin{aligned} p_i(x) &= \prod_{j \in \mathcal{B}_i} \frac{x - s_j}{s_i - s_j} \\ \mathcal{B}_i &= \{j | j \neq i, j \in \mathcal{D}_n\} \end{aligned} \tag{3.22}$$

其中，$\mathcal{D}_n$是参与者的序号集合。当x为0时，$p_i(0)$即为目标扰动，用p_i表示：

$$p_i = p_i(0) = \prod_{j \in \mathcal{B}_i} \frac{-s_j}{s_i - s_j} \tag{3.23}$$

为了进一步防止服务器与参与者之间的共谋，秘密参数$\{s_i\}$不能直接在参与者之间传递。利用Shamir秘密共享（可加同态）和朴素的乘法秘密共享（可乘同态）来私下计算每个参与者i的p_i。具体而言，对于每个参与者i，其首先选择一个本地的掩蔽比例参数η_i来掩盖s_i。然后，参与者i为其他参与者生成s_i的Shamir秘密份额$\{j, s_{i,j}\}$和η_i的乘法秘密份额$\{j, \eta_{i,j}\}$。接下来，秘密份额和掩蔽后的$\eta_i s_i$被传输给相应的参与者。参与者k利用Shamir秘密共享的可加同态特性计算$g_{i,j}^k$：

$$g_{i,j}^k = s_{i,k} - s_{j,k} \tag{3.24}$$

秘密份额$\{g\}$被发送给一个被认为不会与服务器共谋的可信任参与者l。l在接收关于$s_i - s_j$的足够秘密份额后，可以重建原始的秘密差值$s_i - s_j$作为中继节点。然后l计算所有秘密差值的乘积，即$\prod\limits_{j\in\mathcal{B}_i} s_i - s_j$并将其发送给相应的参与者$i$。因此，掩蔽后的$\hat{p}_i$可以由参与者$i$计算：

$$\hat{p}_i = \prod_{j\in\mathcal{B}_i} \frac{-\eta_j s_j}{s_i - s_j} \tag{3.25}$$

为了获得原始扰动p_i，需要移除比例参数的乘积$\prod\limits_{j\in\mathcal{B}_i} \eta_j$。每个参与者$i$利用朴素乘法秘密共享的可乘同态特性计算$\tau_{i,j}$：

$$\tau_{i,j} = \prod_{k\in\mathcal{B}_j} \eta_{k,i} \tag{3.26}$$

$\tau_{i,j}$被发送给相应的参与者。通过这种方式，参与者i可以重建原始值$\prod\limits_{j\in\mathcal{B}_i} \eta_j$。最后，每个参与者$i$可以计算扰动$p_i = \hat{p}_i / \prod\limits_{j\in\mathcal{B}_i} \eta_j$，而不泄露秘密参数$s_i$，从而防止服务器与参与者之间的共谋。

然后，p_i被添加到更新中，以扰动式(3.15)中的原始更新。这些扰动在服务器聚合掩码更新时被消除。

$$\begin{aligned}
\Delta\theta &= \sum_{i=1}^{n} w_i \Delta\theta_i' \\
&= \sum_{i=1}^{n} w_i (\Delta\theta_i + p_i) \\
&= \sum_{i=1}^{n} w_i \Delta\theta_i + \sum_{i=1}^{n} w_i p_i \\
&= \sum_{i=1}^{n} w_i \Delta\theta_i
\end{aligned} \tag{3.27}$$

由于$f(x)$的常数项为0，根据拉格朗日插值的性质，很容易看出$\sum\limits_{i=1}^{n} w_i p_i$这一项也等于0。

在DRL-PBFL中需要解决的一个关键问题是如何优化聚合策略$\mathcal{P}$，以便每个参与者的权重w_i能有效反映其模型和数据质量，同时减轻拜占庭参与者的影响，并增加良性参与者对聚合更新$\Delta\theta$的影响。强化学习是一种有效的通用人工智能方法，根据当前环境状态输出优化后的动作。深度Q网络（Deep Q-Network, DQN）是强化学习中最常用的算法之一，它可以解决高维空间中的问题。然而，由于DQN仅适用于离散动作空间而聚合策略$\mathcal{P}$是连续的，故它不适用于DRL-PBFL。在本案例中，使用DDPG算法来更新聚合策略。DDPG是一种策略-价值算法，可以处理高维连续动作空间，DDPG可以根据全局模型的性能有效地

更新聚合策略。特别地，参与者的非独立同分布数据分布和拜占庭参与者的影响最终会降低全局模型在验证数据集上的性能，而性能是聚合策略优化过程中最关键的因素。现有的基于性能的方案通常依赖于本地更新和全局更新之间的相似性，这在数据非独立同分布场景中不适用。因此，DRL-PBFL在数据非独立同分布场景中比其他方案更能取得令人满意的性能。

DDPG组件使用通用的DDPG架构，包括决策网络$\mu(S|\theta_\mu)$、目标决策网络$\mu'(S|\theta_{\mu'})$、评估网络（Q网络）$Q(S,a|\theta_Q)$、目标评估网络$Q^{'}(S,a|\theta_{Q'})$、随机过程噪声生成器$\mathcal{N}_t$和经验回放缓冲区$\mathcal{R}$。图3-9展示了DDPG组件的架构。S和a分别表示状态和动作，θ表示DDPG组件中网络的参数。将聚合策略的更新过程建模为具有状态空间$\mathbb{S}$、动作空间$\mathbb{A}$、奖励函数R、初始状态分布$P(S_1)$和转移动力学$P(S_{t+1}|S_t,a_t)$的马尔可夫决策过程，下标t表示通信轮数。通信轮数t的状态S_t表示元组$S_t=(l_r,\mathcal{P},\mathcal{H})$，如式(3.18)和式(3.19)所示。动作$a$是一个$n$维向量，其中$n$是参与者的数量。输出的动作$a$可以通过归一化转换为聚合策略$\mathcal{P}$。奖励函数$R$反映状态的变化，并为聚合策略的更新提供启发。轮次$t$的奖励$r_t$由本轮次的全局模型损失$l_r(\theta_g^{'})$、上轮次的全局模型损失$l_r(\theta_g)$和当前最佳模型损失$l_r(\theta_g^{\text{best}})$决定：

$$r_t=(l_r(\theta_g)-l_r(\theta_g'))-(l_r(\theta_g')-l_r(\theta_g^{\text{best}})) \tag{3.28}$$

损失l_r根据式(3.18)计算，最佳模型θ_g^{best}在训练过程中由中心服务器记录，θ_g^{best}表示在验证数据集$\mathcal{D}_r$上具有最低测试损失的当前全局模型。对于初始状态分布$P(S_1)$，$\mathcal{P}$被初始化为$w_1=w_2=\cdots=w_n$和$w_1+w_2+\cdots+w_n=1$，l_r取决于模型参数的随机初始化。对于$P(s_{t+1}|S_t,a_t)$，S_{t+1}的$\mathcal{P}$是a_t的归一化，l_r取决于全局模型的更新方式。

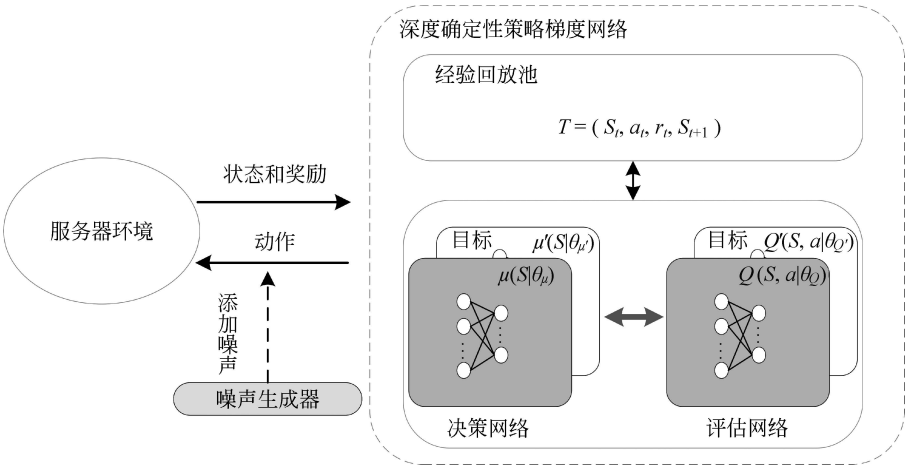

图 3-9　DDPG组件概览

整个DDPG策略更新过程分为两个阶段。第一阶段是填充经验回放缓冲区R。填充过程有两种可能的情况：如果有相同任务和相同参与者的历史信息，可以将信息直接处理成缓冲区所需的数据格式，并填充到缓冲区中；或者存在一个预热期来填充回放缓冲区。对于预热期，动作a_t由决策网络$\mu(S|\theta_\mu)$根据当前状态S_t和探索噪声$\mathcal{N}_t$生成：

$$a_t=\mu(S_t|\theta_\mu)+\mathcal{N}_t \tag{3.29}$$

动作a_t用于更新聚合策略$\mathcal{P}$。然后，观察并计算奖励r_t和下一状态S_{t+1}，分别如式(3.28)

和式(3.18)所示。这样，可得到一个转移元组T：

$$T = (S_t, a_t, r_t, S_{t+1}) \tag{3.30}$$

T存储在经验回放缓冲区$\mathcal{R}$中。探索噪声$\mathcal{N}_t$可以生成随机策略，其中将高权重分配给良性参与者，低权重分配给拜占庭参与者，将生成具有高奖励的经验，反之亦然。这些经验在$\mathcal{R}$中帮助DDPG组件训练神经网络并识别参与者中的拜占庭参与者。当$\mathcal{R}$的大小达到一定阈值时，预热期结束，动作a_t的选择不再涉及探索噪声：

$$a_t = \mu(S_t|\theta_\mu) \tag{3.31}$$

在每个通信轮期间，从$\mathcal{R}$中随机抽取包含N个转移元组(S_i, a_i, r_i, S_{i+1})来更新决策网络和评估网络。由目标评估网络$\theta_Q{}'$输出的元组(S_i, a_i, r_i, S_{i+1})的下一个Q值v_i如下：

$$v_i = r_i + \gamma Q'(S_{i+1}, \mu'(S_{i+1}|\theta_{\mu'})|\theta_{Q'}) \tag{3.32}$$

其中，v_i被称为下一个Q值批次。可以通过最小化Q值批次和下一个Q值批次之间的损失来更新评估网络θ_Q：

$$L = \frac{1}{N}\sum_i (v_i - Q(S_i, a_i|\theta_Q)) \tag{3.33}$$

使用策略梯度算法来更新决策网络μ。μ的策略梯度可以估计为

$$\nabla_{\theta_\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(S_i, a_i|\theta_Q)\nabla_{\theta_\mu}\mu(S_i|\theta_\mu) \tag{3.34}$$

使用 Adam 优化器更新决策网络μ和评估网络Q。最后，软更新目标决策网络μ'和目标评估网络Q'：

$$\begin{aligned} \theta_{Q'} &= \tau\theta_Q + (1-\tau)\theta_{Q'} \\ \theta_{\mu'} &= \tau\theta_\mu + (1-\tau)\theta_\mu \end{aligned} \tag{3.35}$$

其中，τ设置为 0.001。

总体而言，需要多个周期来完成 DDPG 内部神经网络和联邦学习的全局模型θ_g的收敛，一旦θ_g收敛，将得到最终的聚合策略$\mathcal{P}_{\text{final}}$。

3.7 本章小结

本章首先针对无人系统感知安全的现状，从安全威胁、安全案例、安全攻击和对应防御等角度进行了概述，随后，针对无人系统中的传感器攻击、对抗性样本攻击及投毒攻击进行了详细介绍，最后，针对多源数据融合安全，从融合方法、安全融合需求和安全关键技术进行详细介绍。此外，针对基于联邦学习的无人系统中的拜占庭攻击进行了案例分析，具体如下。

（1）无人系统中的传感器包括激光雷达、雷达、摄像头、定位系统、惯性测量单元等。针对这些传感器的攻击可大致分为干扰攻击和欺骗攻击，前者的目的是使传感器失效或降低其性能，后者的目的是操纵传感器所接收的数据，使系统做出错误的判断或反应。

（2）对对抗性样本攻击进行了定义说明，随后根据攻击目标、敌手能力等介绍了对抗性样本攻击的不同分类，并且根据不同的对抗性样本攻击原理介绍了相应的典型对抗性样本

攻击。介绍了对抗性样本攻击的主要防御技术，包括梯度掩蔽/模糊，对抗性防御及对抗性样本检测，并且介绍了相应的实现技术。结合实例说明了对抗性样本攻击在无人系统中的攻击流程和危害性。

（3）对于投毒攻击的基本定义进行了详细说明，随后根据攻击面、攻击目标、攻击者知识、攻击者的方法和能力等介绍了不同分类的投毒攻击，并介绍了典型的投毒攻击实现方法。介绍了多种针对投毒攻击的主动防御和被动防御方法，并结合实例说明了投毒攻击在无人系统中的危害性。

（4）对无人系统中的数据融合技术进行了详细说明，介绍了三类典型数据融合方法。根据无人系统中感知数据的特性，介绍了多源数据融合需求，并且详细说明了多种实现安全可靠多源数据融合所需的关键技术。

（5）介绍了基于深度强化学习的拜占庭鲁棒联邦学习，针对无人系统中的投毒攻击及防御进行了案例分析。

在下一章中，将对无人系统中自主决策层中协同决策、动态攻防对抗、决策可信性评估等进行深入介绍。

3.8 习题

1. 无人系统的感知是指什么？常用的感知传感器有哪些？请列举至少三种传感器，并简要描述它们在典型无人系统中的作用。

2. 什么是无人系统中的数据融合？为什么它对于提高系统的整体感知能力至关重要？

3. 考虑无人驾驶汽车在雾霾天气下进行自动驾驶的情境。讨论在这种特定环境中，数据融合过程中可能面临的挑战，如传感器数据不准确或不完整等。并探讨解决这些挑战的潜在方法。

4. 以城市交通环境为背景，设计一个无人车辆的感知系统。考虑各种环境因素，如交通拥堵、多种天气条件、行人和非机动车辆的存在等，并提出确保系统可靠性的策略。

5. 分析当前的技术趋势，如多模态大模型在感知安全中的应用，并思考这些技术将如何改变未来无人系统的感知安全。