车联网异常检测综述

范仲式

一. 现状

车联网技术已经不断成熟，包括车载通信技术、车载传感器技术、数据处理技术等方面。其中，5G 技术的应用将加速车联网的发展，并提供更快速、更稳定的连接。

通信技术方面，通过 V2X 车载通信手段与 TSP 基础设施，完成人、车、路之间的信息交互与共享。在 V2X 网络中，车辆之间可以相 互交换安全关键信息，如自身传感器探测到的道 路交通状况、交通信号等。此外，通过 V2I 通信技术，车辆可以与附近的基础设施通信，并获得 道路信息、路线规划、交通信息等。

在高精度定位方面，我国主要采用 GNSS，在无 GNSS 信号的地方用其他定位方法作补充。我国正在研究 GNSS 与蜂窝网结合的技术，以期望达到随时随地的定位服务。

除技术方面以外，车联网应用场景已经非常广泛，包括车辆远程监控、驾驶辅助、智能交通、智能停车等。随着技术的不断发展，车联网的应用场景还将不断扩大。市场规模不断扩大，预计到 2025 年，全球车联网市场规模将达到 1.4 万亿美元。在中国，车联网市场也在快速发展，预计到 2025 年，中国车联网市场规模将达到 6000 亿元。政府对车联网的支持力度也在不断加大。例如，中国政府已经发布了多项支持车联网发展的政策，包括资金支持、政策扶持等。此外，车联网的发展还需要各个行业之间的合作。目前，汽车制造商、通信运营商、互联网公司等都在积极参与车联网的发展，加强协作，共同推动车联网的发展。

在国家的支持下，车联网发展迅速，但是，作为一个新兴产业，车联网依然面临着许多挑战：

第一，安全问题亟待解决。考虑到自动驾驶、智能网联面向环境较为复杂，因设计漏洞、人员误操作、网络攻击等造成的数据安全、信息安全、功能安全等问题已经成为制约车联网产业发展的一大挑战，在国外一些国家，该问题甚至已经上升到国家战略。我国虽然已经陆续开展了关于车联网网络安全与数据安全等方面法律法规与标准法规的制定工作，但车联网安全问题仍需相关技术支撑才能真正得以解决。

第二，量产落地亟待推行。《中国互联网发展报告（2021）》数据显示，车联网市场渗透率仅为 15%。目前，仅少量高端 车型搭载车联网相关产品，像一汽红旗 E-HS9、 上汽奥迪 A7L、A6L 等车型搭载前向碰撞预警（Forward Collision Warning, PCW）、交叉路口碰 撞预（Intersection Collision Warning, ICW）等 V2X 技术。此外，车载通信系统目前也无法满足车联网时延以及长距运输等要求。整体来看，车联网端、管、云层均尚未发展成熟，造成车联网装车量较低，制约车联网产品量产落地。

第三，商业模式亟待明晰。从面向对象来说，车联网产品定位与用户实际需求不相契合，用户较难理解并感知产品较强的技术属性；从投资规模来说，车联网产品实现道路全覆盖，需投入较大道路改造资金；从实施单位来说，目前智能网联、自动驾驶尚不明晰交通事故责任，加之我国不同类型的道路基础设施投资、建设和运营主体较多，因此，车联网产品如何

投放等商业方案尚不明确，制约车联网产品推广落地及大规模量产。

二. 异常检测

1. 基于统计的方法

基于统计的方法是通过对数据进行统计分析，寻找偏离正常值的数据点。使用这类方法基于的基本假设是，正常的数据是遵循特定分布形式的，并且占了很大比例，而异常点的位置和正常点相比存在比较大的偏移。常用的统计方法包括：均值-方差方法、箱线图法、Z-score 法等。使用这种方法存在的问题是，均值和方差本身都对异常值很敏感，因此如果数据本身不具备正态性，就不适合使用这种检测方法。

2. 基于深度的方法

基于深度的方法，即从点空间的边缘定位异常点，按照不同程度的需求，决定层数及异常点的个数。

3. 基于密度的方法

该类方法是针对所研究的点，计算它的周围密度和其临近点的周围密度，基于这两个密度值计算出相对密度，作为异常分数。即相对密度越大，异常程度越高。基于的假设是，正常点与其近邻点的密度是相近的，而异常点的密度和周围的点存在较大差异。

4.基于偏差的方法

这是一种比较简单的统计方法，最初是为单维异常检测设计的。给定一个数据集后，对每个点进行检测，如果一个点自身的值与整个集合的指标存在过大的偏差，则该点为异常点。

具体的实现方法是，定义一个指标 SF（Smooth Factor），这个指标的含义就是当把某个点从集合剔除后方差所降低的差值，我们通过设定一个阈值，与这些偏差值进行比较来确定哪些点存在异常。这个方法是由 Arning 在 1996 年首次提出的。

5.基于距离的方法

基于距离的方法，即计算每个点与周围点的距离，来判断一个点是不是存在异常。基于的假设是正常点的周围存在很多个近邻点，而异常点距离周围点的距离都比较远。

6.基于机器学习的方法

基于机器学习的方法是利用机器学习算法对数据进行建模，然后利用模型预测哪些数据点是异常点。

三. 常用车辆异常行为数据集

1. UAH（Utah Automotive Dataset）

UAH 是由美国犹他大学交通工程系和机械工程系合作开发的车辆异常行为数据集。该数据集包括从多个车辆中收集的 GPS 轨迹数据、车速数据、加速度数据等，涵盖了急加速、急刹车、快速转弯等异常行为。该数据集可以用于车辆驾驶行为分析、驾驶员行为评估等方面。

2. NGSIM（Next Generation Simulation）

NGSIM 是由美国联邦公路管理局开发的车辆行驶数据集。该数据集包括在美国加利福尼亚州和华盛顿州的多个高速公路和城市道路上收集的 GPS 轨迹数据、车速数据、加速度数据等。该数据集涵盖了变道、超车、急转

弯等异常行为，可以用于驾驶行为研究、交通流模拟等方面。

3. DAS（Driving Activity Study）

DAS 是由美国国家运输安全委员会开发的车辆异常行为数据集。该数据集包括从多个车辆中收集的 GPS 轨迹数据、车速数据、加速度数据等，涵盖了超速、急加速、急刹车等异常行为。该数据集可以用于驾驶行为分析、交通安全预警等方面。

4. Drive&Act

Drive&Act 是由德国联邦公路研究所开发的车辆异常行为数据集。该数据集包括从多个车辆中收集的 GPS 轨迹数据、车速数据、加速度数据等，涵盖了急转弯、急加速、急刹车等异常行为。该数据集可以用于驾驶行为识别、行驶安全预警等方面。

5. DIDI-MASS

DIDI-MASS 是由中国滴滴公司开发的车辆异常行为数据集。该数据集包括从多个车辆中收集的 GPS 轨迹数据、车速数据、加速度数据等，涵盖了超速、疲劳驾驶、急刹车等异常行为。该数据集可以用于驾驶行为分析、交通安全预警等方面。

四．主流深度学习模型

1.卷积神经网络（Convolutional Neural Network，CNN）

卷积神经网络是一种广泛用于图像处理和计算机视觉的深度学习模型。CNN 通过多层卷积和池化操作来提取图像中的特征，并用全连接层进行分类或回归。常用的 CNN 架构包括 LeNet、AlexNet、VGG、GoogLeNet、ResNet 等。

2.循环神经网络（Recurrent Neural Network，RNN）

循环神经网络是一种用于处理序列数据的深度学习模型。RNN 通过在网络中引入循环结构来处理时序依赖关系，使得网络能够记忆之前的状态并在当前状态中使用。常用的 RNN 架构包括基本的 RNN、长短时记忆网络（LSTM）、门控循环单元（GRU）等。

3.生成对抗网络（Generative Adversarial Network，GAN）

生成对抗网络是一种用于生成新的数据样本的深度学习模型。GAN 由两个神经网络组成，一个生成器网络和一个判别器网络。生成器网络从一个随机噪声中生成新的数据样本，而判别器网络则尝试区分真实数据和生成器生成的数据。通过训练生成器网络来迫使其生成的数据与真实数据分布相似，从而达到生成新数据的目的。

4.注意力机制（Attention Mechanism）

注意力机制是一种用于处理序列数据和图像数据的深度学习模型。注意力机制通过对输入数据中不同位置或特征的重要性进行加权，使得模型能够在处理数据时更加关注重要的信息。常用的注意力机制包括序列到序列（Seq2Seq）模型中的注意力机制和图像分类中的空间注意力机制等。

5.自编码器（Autoencoder）

自编码器是一种用于无监督学习和特征提取的深度学习模型。自编码器通过将输入数据编码为压缩表示，然后再将其解码为原始数据，从而使得模型能够学习到输入数据中的重要特征。常用的自编码器包括标准的自编码器、变分自编码器（VAE）、生成对抗自编码器（GAN-AE）等。

五. 车联网异常检测研究现状

异常行为检测系统是确保 5GB 车载网络安全的关键组成部分。机器学习是设计这些系统不可或缺的工具。目前，用于 5GB 车载网络的基于 ML 的 MDS 仍处于开发的第一阶段。

在车联网异常检测研究的研究过程中，第一步可能是增强对 5GB 环境中针对 CAV 的攻击及其潜在场景的了解。这一步骤对于构建真实的测试平台以生成可靠的攻击数据集至关重要，可以作为验证和比较结果的基准。

下一步是仔细研究基于 ML 的 MDS 的部署，考虑不同的性能指标、ML 模型的安全性和激励函数，以确保可持续性。MLOps 是这个阶段的一个关键功能。MLOps 是一个开发领域，其原理和工具有助于 ML 项目的生命周期，尤其是数据处理、模型构建和部署。在部署基于 ML 的 MDS 时，MLOps 可以同时解决 ML 和软件工程问题。ML 问题主要包括数据和概念漂移。另一方面，MLOps 可以帮助解决软件工程问题。根据检测到的攻击，MLOps 可以帮助决定是部署基于 ML 的 MDS 来进行实时检测还是批量检测，以及部署基于 ML 的 MDS 的最佳位置。此外，MLOps 工具允许监控基于 ML 的 MDS 消耗处理和内存资源的数量。还可以监控其他实时软件工程的性能，如延迟和吞吐量。MLOps 还提供了记录数据以进行分析和审查的服务，并为重新训练基于 ML 的 MDS 提供了更多数据。关于安全和隐私，MLOps 可以帮助根据数据敏感性和监管要求在基于 ML 的 MDS 上定制适当的安全和隐私级别。

最后一步，浓缩标准化活动将有助于加速行业对基于 ML 的 MDS 的采用。更具体地说，需要在当前标准规定的合理性检查的基础上，定义一种检测不当行为的 ML 方法。此外，标准化机构应进一步关注检测合作感知服务中的不当行为。定义标准规范将使基于 ML 的 MDS 在 5GB 车辆网络中创造新的商业机会。

六. 不足之处

1.数据不完整和标注不准确

车联网数据的获取和标注比较困难，数据的质量和数量都受到很大的限制。数据不完整和标注不准确会严重影响异常检测模型的性能和可靠性。

2.模型泛化能力差

车联网异常检测模型的训练数据往往只来自于某些特定场景或条件，这会导致模型的泛化能力较差，难以适应未知场景和条件下的异常检测任务。

3.缺乏有效特征和算法

车联网异常检测中，数据复杂多变，从中提取出有效的特征来进行异常检测十分困难。当前的特征提取方法和算法在处理车联网数据时仍存在一定的局限性，需要进一步改进和优化。

4.缺乏统一标准和评价指标

目前车联网异常检测领域缺乏统一的标准和评价指标，不同研究者使用的数据集和评价方法不一致，难以进行模型性能的比较和评价。

5.隐私和安全问题

车联网数据涉及到个人隐私和车辆安全等方面的问题，这在异常检测研究中也是一个重要的难点。如何在保证数据隐私和车辆安全的前提下，进行有效的异常检测仍需要进一步研究和探索。

七．可能存在的创新点

1.目前车联网异常检测主要使用车载传感器数据进行分析和检测，而忽略了其他数据源的潜在价值。未来可以探索如何将车载传感器数据与其他数据源（如路况、天气、交通状况等）进行融合，提高异常检测的准确性和鲁棒性。

2. 车联网数据具有时变性和非稳态性等特点，需要能够自适应地调整模型参数和学习策略。未来可以探索如何设计自适应学习方法来应对车联网异常检测任务中的时变性和非稳态性问题。