基于改进能量模型的主动域自适应安全性评估方法

刘 畅 [1] 何 潇 [1] 王立敏 [2]

摘 要 复杂动态系统运行过程中的在线安全性评估至关重要且富有挑战性. 构建有效的数据驱动模型需要大量有标注数据, 但这在实际中通常难以获得. 此外, 考虑到系统不同的运行工况, 安全性评估模型应该具有良好的泛化能力. 域自适应 (Domain adaptation, DA) 可以将模型从数据标注丰富的源域迁移到具有不同但相似数据分布的目标域. 然而, 源域中没有出现过的任务相关未知情景会降低模型的性能, 是目前尚未解决的挑战. 主动域自适应通过结合域自适应与主动学习技术, 为解决上述挑战提供了思路. 本文研究目标域存在任务相关未知情景的主动域自适应安全性评估问题, 提出一种基于改进能量模型的主动域自适应方法. 在所提方法中融合分布外检测器, 在此基础上主动选择目标域中具有代表性的无标注样本进行标注, 作为训练数据以提高域自适应模型的性能. 最后, 通过基于轴承数据的案例研究, 验证所提方法的有效性和适用性.

关键词 在线安全性评估, 域自适应, 主动学习, 基于能量的模型

引用格式 刘畅, 何潇, 王立敏. 基于改进能量模型的主动域自适应安全性评估方法. 自动化学报, 2024, **50**(10): 1928−1937

DOI 10.16383/j.aas.c230685 **CSTR** 32138.14.j.aas.c230685

Active Domain Adaptation for Safety Assessment: An Improved Energy-based Model

LIU Chang[1] HE Xiao[1] WANG Li-Min[2]

Abstract Online safety assessment of complex dynamic systems during operation is paramount and challenging. A large amount of labeled data is necessary to construct an effective data-driven model, which is difficult to obtain in practice. Furthermore, the safety assessment model should have a good generalization ability given the varying operation modes. Domain adaptation (DA) can transfer the model trained on a source domain with abundant labeled data to a target domain that has a different but similar data distribution. However, the task-related unknown scenarios that have not appeared in the source domain will degrade the model performance, which remains an unsolved challenge at present. Active domain adaptation provides a potential solution to the aforementioned challenge by combining domain adaptation with active learning techniques. This paper investigates the problem of active domain adaptation for safety assessment, specifically addressing task-related unknown scenarios within the target domain. An active domain adaptation method with the improved energy-based model is proposed, and the out-of-distribution detector is incorporated in the proposed method. On this basis, representative unlabeled samples from the target domain are actively selected for annotation, which are then used as training data to enhance the performance of the domain adaptation model. At last, a case based on the bearing data is studied to demonstrate the effectiveness and applicability of the proposed method.

Key words Online safety assessment, domain adaptation (DA), active learning, energy-based model (EBM)

Citation Liu Chang, He Xiao, Wang Li-Min. Active domain adaptation for safety assessment: An improved energy-based model. *Acta Automatica Sinica*, 2024, **50**(10): 1928−1937

在当今信息化和自动化时代, 在线评估动态系统的运行安全性[1] 十分重要, 尤其是对于诸如化工过程、智能交通、航空航天等领域中的关键应用. 系统安全性是指不对设备、周围环境和人产生危害的能力. 随着动态系统在复杂环境中持续运行, 对系统的安全性评估必须适应性强, 能够迅速检测出潜在的危险状态, 智能地判断系统当前状态可能产生的影响, 给操作员反馈以采取对应措施. 然而, 评估复杂动态系统的运行安全性是极具挑战性的难题. 一方面, 系统本身具有非线性、强耦合等特点, 其机理模型难以获得; 另一方面, 系统运行环境复杂多变, 伴随着强不确定性. 为提高系统安全性, 近年来

收稿日期 2023-11-07 录用日期 2024-03-21
Manuscript received November 7, 2023; accepted March 21, 2024
国家自然科学基金 (62163012, 62473223), 北京市自然科学基金 (L241016) 资助
Supported by National Natural Science Foundation of China (62163012, 62473223) and Beijing Natural Science Foundation (L241016)
本文责任编委 李鸿一
Recommended by Associate Editor LI Hong-Yi
1. 清华大学自动化系 北京 100084 2. 广州大学机械与电气工程学院 广州 510006
1. Department of Automation, Tsinghua University, Beijing 100084 2. School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006

国内外学者对相关问题进行了广泛研究. 例如, 过程检测[2−5] 技术被用来检测系统在运行过程中是否偏离正常情况, 故障诊断[6−10] 方法旨在识别系统发生的故障类型, 容错控制[11−14] 算法用于确保在故障发生时系统仍能够完成指定任务. 相比之下, 对系统在线安全性评估的关注还远远不够.

系统安全性评估的提出可以追溯到 20 世纪初, 当时化学、核工厂和航空航天等行业开始实施安全管理系统, 以降低与其运行相关的风险. 安全性评估的主要目的是预防事故发生或减轻事故后果. 为了在安全与经济成本之间取得平衡, 安全性评估需要回答“系统当前或在特定条件下的安全程度如何”以及“应该采取哪些维护或自救措施”. 文献 [15] 综述了在航天发射系统中运行安全性评估应用的进展与挑战. 考虑到早期系统安全性评估方法都是基于确定性的定量分析, 概率安全性评估[16] (Probability safety assessment, PSA) 被提出用于估计系统发生危险的概率. 此外, 为了处理时间相关的特性, 动态概率安全性评估[17] (Dynamic probability safety assessment, DPSA) 被提出以克服静态方法的局限性. 然而, 现有的大多数系统安全性评估方法都是离线或基于知识的, 不能用于系统运行安全性的实时评估. 由于动态系统状态随时间变化的特点, 系统运行安全性的在线评估应该受到更多关注, 特别是基于状态监测的数据驱动模型的开发.

在线安全性评估是指利用历史数据建立数据驱动模型, 在系统运行时根据实时监测的数据判断系统的安全状态. 文献 [18] 提出一种基于证据推理的在线安全性评估方法, 同时融合了过去、当前和未来的信息. 文献 [19] 提出一种基于证据组交互的安全性评估方法, 融合了专家经验知识. 近年来, 机器学习领域的发展为在线安全性评估提供了便利, 但在实际应用中仍面临着一些挑战. 通常, 机器学习成功的前提是有大量标注数据且数据服从相同的分布, 但是实际系统的运行环境复杂多变且充满不确定性, 限制了机器学习方法在实际系统中的应用. 文献 [20] 提出一种专家知识增强的数据驱动安全性评估方法, 用属性向量描述系统运行状态, 仅用正常运行数据构建属性检测器, 在异常情景零样本下实现安全性的在线评估. 然而, 实际系统的运行模式往往是多变的. 例如, 旋转机械经常在不同的工作条件 (负载和速度) 下运行, 在这种情况下, 特定运行模式下采集的数据可能是无标注的. 因此, 一个需要解决的问题是将一种工况下训练的模型泛化或迁移到另一种工况.

域自适应[21] (Domain adaptation, DA) 是一种可以解决上述问题的迁移学习技术, 主要用于解决将知识从一个域 (源域) 传递到另一个域 (目标域) 的问题, 以提高模型的泛化性能. 其中, 域通常由数据分布、特征空间或其他特征的差异来定义. 域自适应的目标是使得基于一个领域训练的模型在相似的另一个领域上表现良好. 无监督域自适应 (Unsupervised domain adaptation, UDA) 是域自适应任务的一个子集, 它是指在没有任何目标域数据标签的情况下进行知识迁移. 在这种情况下, 模型只能利用源域的有标签数据和目标域的无标签数据来适应目标域的数据分布. 近年来, 无监督域自适应在故障诊断领域的应用已经获得了很多关注. 文献 [22] 对基于无监督深度迁移学习的智能故障诊断方法进行了全面的综述. 文献 [23] 提出一种基于对比学习的域自适应网络, 用于诊断可变工况下的轴承故障. 传统无监督域自适应方法都是基于封闭世界的假设, 即, 假设源域数据涵盖了所有可能的情景, 并且源域和目标域的类别空间完全相同. 但是, 实际中目标域可能会存在未知情景. 因此, 开放集域自适应问题[24−26] 引起了学术界的关注, 其目标是在模型迁移过程中保证已知类别的分类性能, 同时识别和拒绝未知类别. 文献 [27] 提出实例级加权对抗学习方法, 根据和源域的相似性对目标域样本进行加权, 并基于权重生成目标域未知类别的伪标签, 在此基础上训练二分类器进行未知类别的检测. 文献 [28] 提出自适应开放集域泛化网络, 集成了度量学习来增强特征表示, 同时通过学习表示空间中的类决策边界来引入异常检测模块. 然而, 这些研究都假设目标域中的未知情景属于新颖类别, 将它们视为一个超类且与任务无关. 现有方法不能处理未知情景属于已知类别的情况, 但这种情况在安全性评估任务中很常见. 这是由于不同的故障情景对系统安全的影响程度可能相同, 因此会被划分为相同的安全状态. 考虑到存在任务相关的未知情景, 在线安全性评估模型需要做出及时和正确的处理. 文献 [29] 提出一种基于主动增量学习的安全性评估方法, 可以通过“人在回路”的主动标注实现模型增量更新的闭环, 在任务相关未知情景意外出现时仍能准确评估系统的安全状态. 针对在线安全性评估领域, 任务相关未知情景带来的挑战尚未在域自适应范式下得到充分研究, 是具有创新性的难题.

主动域自适应[30] 是解决上述挑战的一种可行方案. 它结合了域自适应与主动学习的概念, 旨在选择具有代表性的目标域样本进行标签查询, 从而使模型在迁移过程中更有效地适应目标域, 同时尽可能降低标签标注的成本. 相比于无监督域自适应仅利用目标域的无标注数据, 主动域自适应通过设计有效的主动样本标注策略, 能够在有限标注成本

内学习到目标域的知识. 文献 [31] 针对主动域自适应提出一种名为聚类不确定性加权嵌入的新标签获取方法. 文献 [32] 提出一种基于能量模型的主动域自适应方法, 设计一种基于自由能和不确定性的主动样本标注方法. 基于能量的模型[33] (Energy-based model, EBM) 比一般的分类模型具有更好的鲁棒性和分布外样本检测能力[34], 并且其中隐藏的自由能函数可以作为目标域样本出现的合理性度量. 然而, 主动域自适应的研究目前都集中于图像识别领域, 在安全性评估领域的应用尚无研究. 同时, 现有的主动域自适应方法都基于封闭世界的假设, 没有考虑目标域中存在任务相关未知情景的情况, 在给定标注预算下对模型性能的提升效果受限.

综上所述, 本文提出一种新的用于安全性评估的主动域自适应解决方案. 与现有工作相比, 本文的主要贡献总结如下: 1) 首次提出目标域含任务相关未知情景的域自适应安全性评估问题; 2) 提出一种基于改进能量模型的主动域自适应方法, 融合分布外检测器以选择代表性样本进行主动标注; 3) 提供基于轴承数据的案例研究来说明所提方法的有效性和适用性.

1 问题描述

本节给出考虑任务相关未知情景的主动域自适应安全性评估问题的描述. 在此之前, 首先给出相关概念的定义.

定义 1. 已知情景: 在初始时有相关标注数据的系统运行状态, 记作 $\mathcal{S}_K = \{s_1, s_2, \cdots, s_k\}$.

定义 2. 未知情景: 在初始时没有相关标注数据的系统运行状态, 记作 $\mathcal{S}_U = \{s_{k+1}, s_{k+2}, \cdots, s_{k+u}\}$.

假设系统的安全状态可以被分为 m 个等级, 通常表示为 $\mathcal{Y} = \{y_1, y_2, \cdots, y_m\}$. 由于系统安全状态的划分一般都是已知且完备的, 因此本文假设已知情景和未知情景共享相同的安全状态标签空间 $\mathcal{Y}$. 在考虑含有任务相关未知情景的主动域自适应安全性评估任务中, 总体目标是根据源域知识训练针对于目标域的安全性评估模型, 同时主动选择有代表性的目标域样本进行标注, 以适应仅存在于目标域的未知情景. 初始时刻的训练数据集包含了来自源域的有标注数据集 $\mathcal{D}_S = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{n_s}$ 和目标域的无标注数据集 $\mathcal{D}_T = \{\boldsymbol{x}_i^t\}_{i=1}^{n_t}$. 其中, $\boldsymbol{x}_i^*, * \in \{s, t\}$ 是测量数据或其特征表示, y_i^s 是 $\boldsymbol{x}_i^s$ 对应的标签, n_s 和 n_t 分别表示源域数据集和目标域数据集的样本数量. 源域数据集 $\mathcal{D}_S$ 仅来自于已知情景 $\mathcal{S}_K$, 而目标域数据集 $\mathcal{D}_T$ 除了已知情景外, 还有可能来自于未知情景 $\mathcal{S}_U$. 最终的目标是训练一个安全状态分类模型, 能够准确评估目标域测试样本的安全等级. 由于未知情景在初始时没有相关的标注数据, 因此本文采用主动域自适应方法, 在训练过程中选择和标注 $\mathcal{D}_T$ 中的代表性样本, 以此来逐步学习关于未知情景的知识.

2 基于改进能量模型的主动域自适应方法

针对安全性评估模型迁移过程中目标域存在任务相关未知情景的问题, 本文提出一种基于改进能量模型的主动域自适应方法. 该方法的总体框架如图 1 所示, 包含 2 个主要组成部分: 1) 基于能量的模型, 根据有标注数据和无标注数据训练的安全状态分类模型, 能够同时处理目标域的已知情景和未知情景; 2) 主动标注过程, 根据设定的策略选择具有代表性的目标域无标注样本进行主动标注, 将新标注的数据加入到训练数据中.

2.1 基于能量的模型

基于能量的模型[33] 能够通过为变量 $\boldsymbol{x}$ 和 y 的每个组合分配一个标量值作为能量的度量, 来捕获变量之间的依赖关系. 模型训练过程的本质是建立一个能量函数 $E(\cdot)$, 将低能量值与正确的预测联系起来, 将较高的能量值与错误的预测值联系起来. 一般通过具有特定损失函数的神经网络来实现模型. 具体来说, 给定输入实例 $\boldsymbol{x}$, 神经网络模型必须输出期望值 y^*, 使得能量函数 $E(\boldsymbol{x}, y)$ 的值最小化

$$y^* = \arg\min_{y \in \mathcal{Y}} E(\boldsymbol{x}, y) \tag{1}$$

由于能量函数未经校准, 因此它的取值范围不确定. 一般可以采用 Gibbs 分布将所有可能的能量函数输出值转换为归一化的概率分布. 在此基础上, 联合概率密度函数 $p(\boldsymbol{x}, y)$ 可以被表示为

$$p(\boldsymbol{x}, y) = \frac{\exp(-E(\boldsymbol{x}, y))}{\sum\limits_{\boldsymbol{x} \in \mathcal{X}} \sum\limits_{y \in \mathcal{Y}} \exp(-E(\boldsymbol{x}, y))} \tag{2}$$

其中, $\mathcal{X}$ 和 $\mathcal{Y}$ 分别表示 $\boldsymbol{x}$ 和 y 的取值空间.

概率密度函数 $p(\boldsymbol{x})$ 能够指示每个目标域样本出现的可能性, 具体可以通过下式进行估计

$$p(\boldsymbol{x}) = \sum_{y \in \mathcal{Y}} p(\boldsymbol{x}, y) = \frac{\sum\limits_{y \in \mathcal{Y}} \exp(-E(\boldsymbol{x}, y))}{\sum\limits_{\boldsymbol{x} \in \mathcal{X}} \sum\limits_{y \in \mathcal{Y}} \exp(-E(\boldsymbol{x}, y))} \tag{3}$$

然而, 上式中的分母无法直接计算或估计. 因

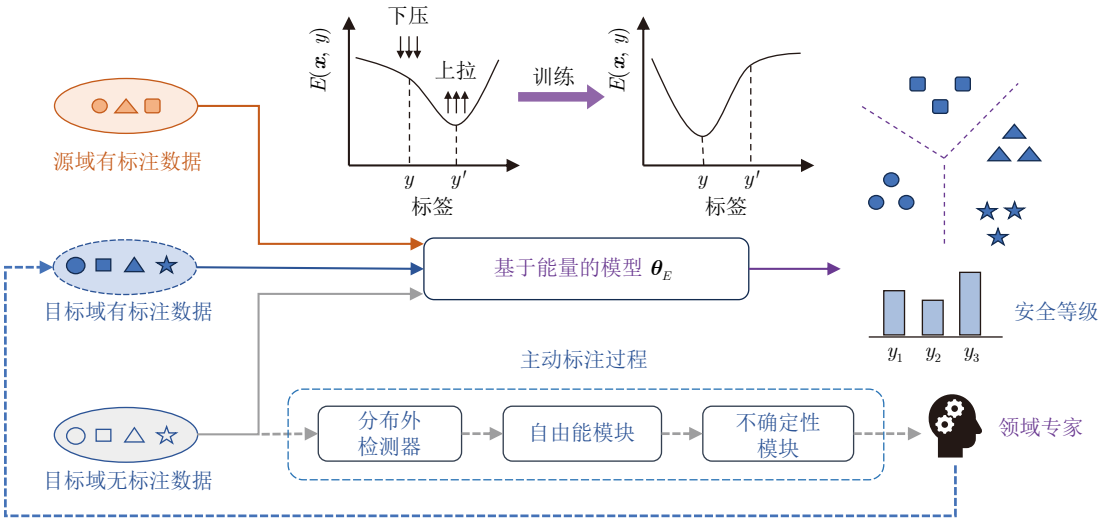

图 1　基于改进能量模型的主动域自适应安全性评估框架

Fig. 1　The safety assessment framework of the active domain adaptation with the improved energy-based model

此, 引入自由能 $\mathcal{F}(\boldsymbol{x})$ 的概念, 用于评估样本 $\boldsymbol{x}$ 出现的合理性, 其表达式如下

$$\mathcal{F}(\boldsymbol{x}) = -\ln \sum_{y\in\mathcal{Y}} \exp(-E(\boldsymbol{x}, y)) \tag{4}$$

通过将式 (4) 替换到式 (3) 中, $p(\boldsymbol{x})$ 可以被重新表示为

$$p(\boldsymbol{x}) = \frac{\exp(-\mathcal{F}(\boldsymbol{x}))}{\sum\limits_{\boldsymbol{x}\in\mathcal{X}} \exp(-\mathcal{F}(\boldsymbol{x}))} \tag{5}$$

文献 [32] 研究表明, 由于源域与目标域之间的数据分布存在偏差, 来自不同域的数据样本会在仅基于源域有标注数据训练的模型上表现出自由能偏差. 具体而言, 大多数源域有标注数据的自由能会低于目标域无标注数据的自由能. 这种自由能存在偏差的特性可以用于选择代表性目标域样本, 以及帮助模型在不同域之间进行泛化. 在此基础上, 文献 [32] 建立了基于能量的域自适应模型, 包含负对数似然损失函数 $\mathcal{L}_1$ 和自由能对齐损失函数 $\mathcal{L}_2$. 其中, 负对数似然损失函数 $\mathcal{L}_1$ 旨在根据有标注数据训练能量函数的参数, 其表达式如下

$$\mathcal{L}_1(\boldsymbol{x}, y; \boldsymbol{\theta}_E) = E(\boldsymbol{x}, y) + \frac{1}{\tau}\ln \sum_{c\in\mathcal{Y}} \exp(-\tau E(\boldsymbol{x}, c)) \tag{6}$$

其中, $\boldsymbol{\theta}_E$ 是能量函数的参数, τ 是逆温度参数. 在本文实验中, τ 设置为 1. 式 (6) 等号右边的第一项用于拉低期望输出的能量值, 而第二项用于推高所有输出的能量值, 以此确保期望输出的能量值最小.

自由能对齐损失函数 $\mathcal{L}_2$ 作为正则项, 利用目标域的无标注数据减小不同域之间的自由能偏差, 以此来增强模型在不同域之间的泛化能力. 具体表达式如下

$$\mathcal{L}_2(\boldsymbol{x}; \boldsymbol{\theta}_E) = \max(0,\ \mathcal{F}(\boldsymbol{x}) - \Delta) \tag{7}$$

其中, $\Delta = \mathrm{E}_{\boldsymbol{x}\sim\mathcal{D}_S}\mathcal{F}(\boldsymbol{x})$ 是根据源域有标注数据计算得到的平均自由能, E 表示数学期望. 在实验中, Δ 与文献 [32] 中一样通过指数滑动平均进行估计. 式 (7) 可以约束目标域样本的自由能小于 Δ, 从而减小源域与目标域的自由能偏差, 实现目标域和源域的分布对齐.

主动域自适应的目标是在训练过程中选择和标注具有代表性的目标域样本, 将其加入到下一轮训练过程中以提升模型性能. 因此, 目标域训练数据集可以分为有标注数据集 $\mathcal{D}_T^l = \{\boldsymbol{x}_i^t, y_i\}_{i=1}^{n_{t1}}$ 和无标注数据集 $\mathcal{D}_T^u = \{\boldsymbol{x}_i^t\}_{i=1}^{n_{t2}}$. 综上, 基于能量的模型[32] 构建的总体目标损失函数如下

$$\min_{\boldsymbol{\theta}_E} \mathrm{E}_{(\boldsymbol{x},\, y)\sim\mathcal{D}_S\cup\mathcal{D}_T^l}\mathcal{L}_1(\boldsymbol{x},\ y; \boldsymbol{\theta}_E) + \gamma \mathrm{E}_{\boldsymbol{x}\sim\mathcal{D}_T^u}\mathcal{L}_2(\boldsymbol{x}; \boldsymbol{\theta}_E) \tag{8}$$

其中, γ 是超参数, 用于平衡 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 的权重.

能量函数的参数 $\boldsymbol{\theta}_E$ 可以用梯度下降的方式进行更新

$$\boldsymbol{\theta}_E := \boldsymbol{\theta}_E - \eta\left(\frac{\partial}{\partial \boldsymbol{\theta}_E}\mathcal{L}_1(\boldsymbol{x},\ y; \boldsymbol{\theta}_E) + \gamma\frac{\partial}{\partial \boldsymbol{\theta}_E}\mathcal{L}_2(\boldsymbol{x}; \boldsymbol{\theta}_E)\right) \tag{9}$$

式中, η 是学习速率. 在实际中可以用 Adam 等优化器对参数 $\boldsymbol{\theta}_E$ 进行求解.

2.2　主动标注过程

如前所述, 主动域自适应的关键问题是如何从目标域无标注数据集中选择具有代表性的样本进行

主动标注. 文献 [32] 提出利用自由能偏差和样本不确定性选择具有代表性的目标域样本进行主动标注. 域自适应模型有效的前提条件是源域和目标域的数据分布相似, 这意味着域之间的分布差异越小, 模型迁移的效果越好. 但是在本文研究的问题中, 目标域中存在未知情景, 由此导致的分布不匹配会对模型产生负面影响. 在这种情况下, 本文提出引入分布外检测器, 根据目标域样本的分布不匹配程度对其价值进行排序.

分布外检测器用于识别与正常数据分布不一致的样本. 令分布外检测器记作 $\mathcal{G}$, 其输入为数据样本, 输出为数据样本的分布外检测得分, 得分越高表明数据的分布不匹配程度越大. 模型 $\mathcal{G}$ 的训练是基于分布内的数据, 在本文中将源域有标注数据 $\mathcal{D}_S$ 作为训练数据. 一般而言, 由未知情景导致的分布不匹配比域之间的固有分布差异更加显著. 因此, 目标域中检测出的分布外样本更有可能来自于未知情景. 针对目标域无标注数据样本 $\boldsymbol{x}_i^t$, 其分布外检测得分可以表示为

$$s(\boldsymbol{x}_i^t) = \mathcal{G}^*(\boldsymbol{x}_i^t) \tag{10}$$

其中, $\mathcal{G}^*$ 是基于数据集 $\mathcal{D}_S$ 训练得到的模型, 相关模型参数记作 $\boldsymbol{\theta}_{\mathcal{G}}^*$.

在分布外检测或异常检测任务中, 一般通过设定合理的阈值 δ, 然后基于下述逻辑判断识别出分布外样本或异常样本

$$\boldsymbol{x}_i^t = \begin{cases} \text{分布内 (正常)}, & s(\boldsymbol{x}_i^t) < \delta \\ \text{分布外 (异常)}, & s(\boldsymbol{x}_i^t) \geq \delta \end{cases} \tag{11}$$

然而, 如何选择合适的阈值 δ 是需要考虑的重要因素. 不同于式 (11) 的逻辑判断形式, 本文对测试数据的分布外检测得分 $s(\boldsymbol{x}_i^t)$ 进行排序, 优先选择分布不匹配程度大的样本, 这样有助于模型学习目标域未知情景的知识, 使得模型针对目标域未知情景的性能表现更好. 值得注意的是, 分布外检测算法的选择并不固定, 可以根据数据集选择有效的算法. 在本文的实验中, 选取一些典型的分布外检测算法进行对比, 包括 COPOD[35]、ECOD[36]、IForest[37]、KDE[38]、OCSVM[39] 和 LOF[40]. 如何设计有效的分布外检测算法是未来值得进一步研究的方向. 除了分布外检测器之外, 本文还考虑了自由能偏差和样本不确定性[32] 作为样本主动标注选择的策略, 如图 2 所示.

具体而言, 样本主动标注的流程如下:

步骤 1. 根据源域数据训练分布外检测器 $\mathcal{G}^*$, 基于 $\mathcal{G}^*$ 计算目标域无标注数据集 $\mathcal{D}_T^u$ 中每个样本的分布外检测得分 $s(\boldsymbol{x}_i^t)$, 并按照降序排列. 选择最

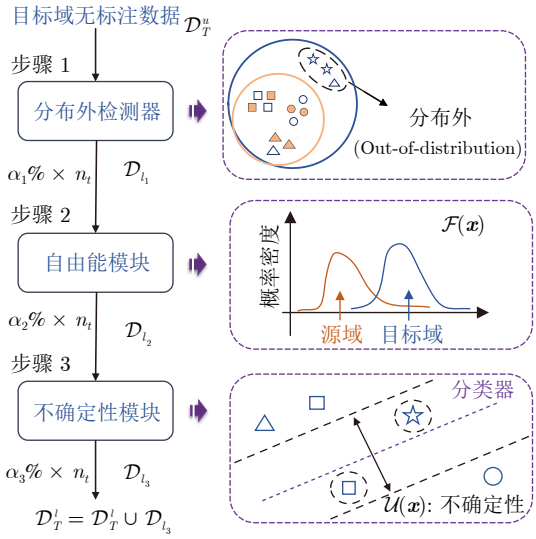

图 2 主动标注过程

Fig. 2 Active labeling process

靠前的 $\alpha_1\% \times n_t$ 个样本组成数据集 $\mathcal{D}_{l_1}$. 分布外检测器仅基于源域数据训练得到, 独立于域自适应训练过程. 相比之下, 自由能偏差策略在模型训练过程中, 随着自由能逐渐对齐可能会对目标域中的分布外样本不敏感.

步骤 2. 具有较高自由能的目标域样本意味着与源域的差异性更大, 所以更能反映目标域的数据分布特点. 因此, 基于当前能量模型的参数 $\boldsymbol{\theta}_E$, 计算步骤 1 得到的数据集 $\mathcal{D}_{l_1}$ 中每个样本的自由能 $\mathcal{F}(\boldsymbol{x}_i^t)$, 并按照降序排列. 选择最靠前的 $\alpha_2\% \times n_t$ 个样本组成数据集 $\mathcal{D}_{l_2}$.

步骤 3. 模型关于样本的不确定性通常可以用来反映模型当前预测的可信程度. 采用能量函数输出的 MvSM (Min-versus-second-min) 值来衡量当前模型下样本的不确定性, 如下式所示

$$\mathcal{U}(\boldsymbol{x}_i^t) = E(\boldsymbol{x}_i^t,\, y^*) - E(\boldsymbol{x}_i^t,\, y') \tag{12}$$

其中, $\mathcal{U}(\boldsymbol{x}_i^t)$ 表示数据样本 $\boldsymbol{x}_i^t$ 的不确定性, $y^* = \arg\min_{y\in\mathcal{Y}} E(\boldsymbol{x}_i^t,\, y)$ 表示具有最低能量函数值的输出标签, $y' = \arg\min_{y\in\mathcal{Y}\setminus\{y^*\}} E(\boldsymbol{x}_i^t,\, y)$ 表示具有次低能量函数值的输出标签. 根据式 (12), 样本 $\boldsymbol{x}_i^t$ 越靠近模型的决策边界, 模型关于样本预测结果的置信度越低, $\mathcal{U}(\boldsymbol{x}_i^t)$ 的值越大.

因此, 计算步骤 2 得到的数据集 $\mathcal{D}_{l_2}$ 中每个样本的不确定性 $\mathcal{U}(\boldsymbol{x}_i^t)$, 并按照降序排列. 选择最靠前的 $\alpha_3\% \times n_t$ 个样本组成最终需要主动标注的数据集 $\mathcal{D}_{l_3}$.

综上, 本文提出的主动域自适应安全性评估方法的实施流程总结为算法 1.

算法 1. 基于改进能量模型的主动域自适应安全性评估

输入. 有标注源域数据集 $\mathcal{D}_S$, 无标注目标域数据集 $\mathcal{D}_T^u = \mathcal{D}_T$, 有标注目标域数据集 $\mathcal{D}_T^l = \emptyset$, 训练轮数 T, 主动标注轮次集合 Q, 超参数 γ, α_1, α_2, α_3, 测试数据 $\boldsymbol{x}_{\text{test}}$.

输出. 测试数据的安全等级标签 y_{test}.

训练阶段

1) 基于 $\mathcal{D}_S$ 训练分布外检测器 $\mathcal{G}^*$;

2) **for** $i = 1 : T$ **do**

3)　根据式 (9) 更新能量模型参数 $\boldsymbol{\theta}_E$;

4)　**if** $i \in Q$ **then**

5)　根据式 (10) 计算样本 $\boldsymbol{x}_i^t \in \mathcal{D}_T^u$ 的分布外检测得分 $\mathcal{G}^*(\boldsymbol{x}_i^t)$;

6)　选取数据集 $\mathcal{D}_T^u$ 中前 $\alpha_1\% \times n_t$ 个具有较高分布外检测得分 $\mathcal{G}^*(\boldsymbol{x}_i^t)$ 的数据样本, 构成数据集 $\mathcal{D}_{l_1}$;

7)　根据式 (4) 计算样本 $\boldsymbol{x}_i^t \in \mathcal{D}_{l_1}$ 的自由能 $\mathcal{F}(\boldsymbol{x}_i^t)$;

8)　选取数据集 $\mathcal{D}_{l_1}$ 中前 $\alpha_2\% \times n_t$ 个具有较高自由能 $\mathcal{F}(\boldsymbol{x}_i^t)$ 的数据样本, 构成数据集 $\mathcal{D}_{l_2}$;

9)　根据式 (12) 计算样本 $\boldsymbol{x}_i^t \in \mathcal{D}_{l_2}$ 的不确定性 $\mathcal{U}(\boldsymbol{x}_i^t)$;

10)　选取数据集 $\mathcal{D}_{l_2}$ 中前 $\alpha_3\% \times n_t$ 个具有较高不确定性 $\mathcal{U}(\boldsymbol{x}_i^t)$ 的数据样本, 构成数据集 $\mathcal{D}_{l_3}$;

11)　更新目标域有标注数据集和无标注数据集:
$\mathcal{D}_T^l \leftarrow \mathcal{D}_T^l \cup \mathcal{D}_{l_3}$, $\mathcal{D}_T^u \leftarrow \mathcal{D}_T^u \setminus \mathcal{D}_{l_3}$

12)　**end if**

13) **end for**

测试阶段

14) 根据式 (1) 评估测试数据的安全等级标签 y_{test};

15) return y_{test}.

3　实验及分析

3.1　实验数据

为了说明所提方法的有效性, 本文使用凯斯西储大学轴承数据集进行实验研究. 凯斯西储大学轴承数据集是从图 3 所示的滚动轴承故障测试平台采集的来自加速度传感器的振动信号数据, 包含不同负载下的正常情景数据和各种故障情景数据, 目前在故障诊断领域被广泛应用于实验研究.

本实验中, 使用的是驱动端采集的加速度传感器数据, 其采样频率为 12 kHz. 实验共考虑三种不同负载下的运行工况, 包括 1 HP、2 HP 和 3 HP, 对应的电机转速分别为 1 772 rpm、1 750 rpm 和 1 730 rpm. 在每种运行工况下, 分别考虑 10 种情景, 包括正常情景以及发生在不同位置 (内圈 IR、滚动体 B 和外圈 OR) 和对应不同严重程度 (7 mils、14 mils 和 21 mils) 的故障情景. 与故障诊断任务旨在准确识别发生的具体故障类型不同, 安全性评估旨在评估故障发生的严重程度及其对系统安全性的影响. 在本实验中, 根据故障严重程度划分对应的安全等级, 如表 1 所示. 考虑到实际系统中发生在某些位置的故障或者较轻微的故障对系统安全性几乎没有影响, 因此实验中将故障大小为 7 mils 的故障情景和正常情景划分为相同的安全等级.

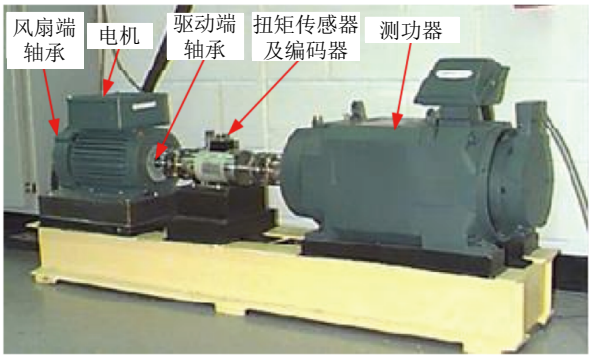

图 3　凯斯西储大学滚动轴承故障测试平台

Fig. 3　Rolling bearing fault test platform from Case Western Reserve University

表 1　不同情景及其对应的安全等级标签

Table 1　Different scenarios and corresponding safety level label

情景	安全等级标签
正常	1
IR-7	1
B-7	1
OR-7	1
IR-14	2
B-14	2
OR-14	2
IR-21	3
B-21	3
OR-21	3

注: x-y 表示发生在位置 x 、故障直径为 y 的故障情景. 例如, IR-7 表示发生在内圈、故障直径为 7 mils 的故障情景.

3.2　实验设置

为了评估方法在目标域存在未知情景时的性能, 实验中将表 1 所列的情景划分为已知情景和未知情景. 其中, 未知情景仅存在于目标域中, 而已知情景则同时存在于源域和目标域中. 具体而言, 本文分别将发生在不同位置的故障设置为未知情景, 而其余情景设置为已知情景.

在实验中选取 3 种方法与本文提出的方法进行

对比, 包括:

1) 基于能量的域自适应模型 (EDA): 以式 (8) 为目标损失函数的模型, 其中目标域数据均无标注, 即 $\mathcal{D}_T^u = \mathcal{D}_T$, $\mathcal{D}_T^l = \emptyset$;

2) 基于能量的主动域自适应模型 (EADA): 文献 [32] 提出的主动域自适应模型;

3) 随机选择策略下的主动域自适应模型 (RAND): 以式 (8) 为目标损失函数的模型, 随机选择给定预算下的目标域无标注数据进行主动标注.

为了衡量不同方法的性能, 在实验中每组实验随机运行多次, 使用平均准确率 ACC 和准确率标准差 STD 作为模型性能评价指标. 它们的计算方式如下式所示

$$ACC = \frac{\sum_{i=1}^{N} ACC_i}{N} \tag{13}$$

$$ACC_i = \frac{|\,C\,|}{|\,T\,|} \tag{14}$$

$$STD = \sqrt{\frac{\sum_{i=1}^{N}(ACC_i - ACC)^2}{N-1}} \tag{15}$$

其中, N 是随机运行的次数, 在本实验中设置 $N = 3$; $|\,C\,|$ 是测试数据集被正确分类的样本数; $|\,T\,|$ 是测试数据集的所有样本数. 显然, 平均准确率越高, 准确率标准差越低, 说明方法的性能越好.

原始一维振动信号被分割为具有 1 024 个采样点的片段, 以此作为模型输入. 所有模型的主干网络如图 4 所示, 均包含 4 个一维卷积层 (Conv) 和 2 个全连接层 (FC). 每个 Conv 层后面都连接了批量归一化层 (BN) 和 ReLU 激活函数. 此外, 第 2 个 Conv 层后面还连接了最大池化层 (Max-pool), 而第 4 个 Conv 层后面还连接了自适应池化层 (AdaptiveMax-pool). 最后, 两个 FC 层之间连接了 ReLU 激活函数、Dorpout 层和 BN 层.

本实验中分别将三种运行工况 (1 HP、2 HP 和 3 HP) 设置为源域和目标域, 有标注的源域数据由来自源域的已知情景产生, 无标注的目标域数据由来自目标域的已知和未知情景产生. 此外, 测试数据由目标域的所有情景产生. 训练集数据和测试集数据的划分比例为 4:1, 参数 γ 设置为 0.01, α_1, α_2, α_3 分别设置为 90、50 和 1, 以确保每次选择 1% 的无标注数据进行主动标注. 对于所有主动域自适应方法, 标签预算设置均相同. 模型训练轮数 T 设置为 50, 主动标注轮次集合 Q 设置为 [10, 15, 20, 25, 30], 意味着训练过程中一共选择 5% 的无标注数据进行主动标注.

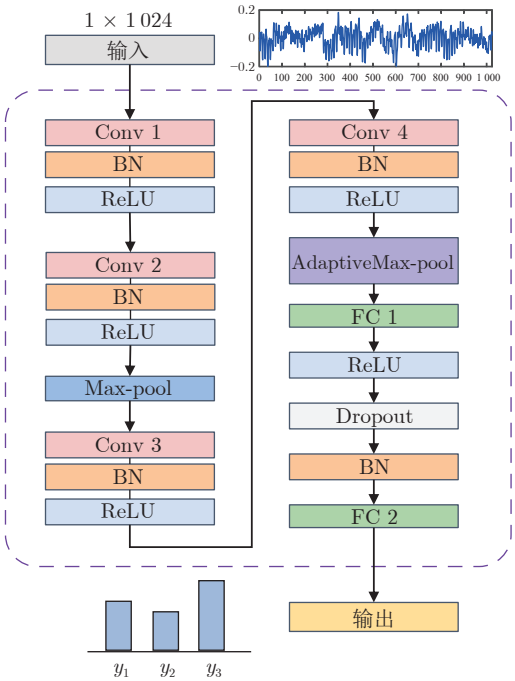

图 4 模型的主干网络

Fig. 4 Backbone network of the model

3.3 实验结果与分析

考虑分别将三种不同的运行工况设置为源域和目标域, 并将发生在不同位置的故障设置为未知情景, 一共有 18 组不同的实验设定. 各种方法在不同实验设定下的安全性评估结果如表 2 所示. 其中, 每组实验中平均准确率最高的方法已加粗显示. 从表 2 可以看出, 主动域自适应方法的表现在大多数情况下都优于传统的域自适应方法, 并且本文所提方法在大多数情况下都优于对比方法. 然而, 在个别情况下, 主动域自适应方法的效果不如传统域自适应方法. 对这一现象的解释是有时未知情景恰好被模型正确地分类. 但尽管如此, 由于没有学习过关于未知情景的知识, 传统域自适应模型在这种情况下仍然是不可靠的.

此外, 为了衡量方法的整体性能, 计算各方法在不同未知情景设定下的平均评估准确率 $\overline{ACC}$, 结果在表 3 中给出. 可以看出, 所提方法的平均性能在大多数情况下都优于对比方法, 并且不同分布外检测算法对模型性能的改善程度不同. 其中, IForest 算法表现最好, 在内圈故障 (IR) 和滚动体故障 (B) 分别设置为未知情景时, 它在不同迁移任务下的平均性能最优.

最后, 不同方法在所有任务上的整体平均准确率在图 5 中给出. 可以观察到, 在主动域自适应方法中, 只有随机选择策略导致了性能下降, 而其他

表 2　不同方法在各种任务和不同未知情景下的安全性评估准确率 $ACC\uparrow(STD\downarrow)$ 比较

Table 2　The safety assessment accuracy $ACC\uparrow(STD\downarrow)$ comparison of different methods under various tasks and different unknown scenarios

任务设定	未知情景	EDA	EADA	RAND	本文方法 (COPOD)	本文方法 (ECOD)	本文方法 (IForest)	本文方法 (KDE)	本文方法 (OCSVM)	本文方法 (LOF)
	IR	77.06 (1.73)	80.66 (3.90)	76.55 (2.54)	76.05 (1.32)	77.34 (1.02)	**81.67** (1.95)	79.44 (4.04)	77.63 (4.42)	79.00 (4.50)
1hp → 2hp	B	**83.62** (0.54)	81.96 (2.75)	**83.62** (0.25)	83.04 (1.54)	81.10 (2.62)	82.47 (3.58)	81.82 (3.94)	81.39 (3.97)	82.76 (0.76)
	OR	**82.40** (0.54)	82.32 (0.12)	80.30 (0.87)	80.88 (1.27)	81.17 (0.37)	81.67 (1.52)	81.02 (0.87)	79.94 (2.62)	82.11 (0.82)
	IR	74.86 (2.20)	72.13 (2.28)	75.36 (1.39)	74.50 (0.12)	74.21 (0.66)	75.36 (3.46)	74.64 (2.99)	**77.37** (2.85)	74.14 (3.56)
1hp → 3hp	B	81.75 (4.38)	81.18 (3.48)	77.95 (1.11)	81.68 (1.41)	**82.54** (1.35)	82.26 (0.90)	80.46 (1.86)	81.47 (3.26)	82.18 (0.90)
	OR	77.16 (1.71)	79.17 (0.82)	**79.74** (2.75)	77.08 (3.99)	78.52 (3.14)	77.30 (0.69)	78.74 (1.59)	79.53 (3.17)	78.59 (1.59)
	IR	70.92 (1.32)	73.52 (1.19)	72.37 (2.05)	74.75 (1.44)	73.09 (1.02)	73.67 (1.62)	72.58 (1.30)	**77.29** (2.06)	74.03 (0.57)
2hp → 1hp	B	84.49 (0.70)	83.98 (1.08)	83.62 (1.02)	84.20 (0.57)	**84.85** (0.57)	83.91 (0.25)	83.91 (0.45)	84.20 (0.43)	84.49 (0.54)
	OR	79.15 (0.90)	79.22 (1.35)	**79.65** (1.77)	79.15 (0.98)	79.22 (0.87)	79.43 (1.21)	79.37 (1.64)	78.35 (1.35)	79.29 (0.76)
	IR	75.22 (1.08)	73.35 (3.42)	73.99 (0.87)	75.07 (2.09)	76.08 (0.57)	76.29 (0.75)	**77.01** (0.90)	75.86 (0.78)	74.93 (0.45)
2hp → 3hp	B	83.48 (1.73)	83.91 (1.08)	83.62 (0.22)	83.76 (0.12)	82.90 (1.32)	82.97 (3.02)	**84.12** (0.69)	83.05 (1.40)	82.97 (1.20)
	OR	82.47 (1.65)	80.68 (1.79)	80.24 (2.24)	82.26 (0.66)	81.25 (0.78)	80.53 (1.74)	80.82 (1.71)	**82.83** (0.82)	81.32 (1.53)
	IR	72.08 (2.13)	73.16 (0.99)	71.43 (0.94)	72.44 (0.62)	**75.11** (2.81)	71.57 (0.66)	72.08 (2.34)	72.73 (0.43)	73.38 (0.99)
3hp → 1hp	B	83.04 (0.33)	84.63 (0.78)	83.98 (1.32)	84.13 (0.45)	84.56 (0.54)	**85.93** (0.37)	85.35 (1.32)	84.34 (0.45)	84.70 (1.75)
	OR	72.29 (2.38)	71.00 (1.35)	69.77 (0.87)	73.02 (2.25)	71.28 (1.84)	70.09 (2.76)	**74.96** (1.84)	74.24 (1.52)	70.63 (1.44)
	IR	76.70 (1.89)	76.26 (0.76)	77.34 (1.68)	76.84 (0.94)	76.41 (1.42)	**78.14** (1.52)	76.41 (1.85)	75.61 (2.14)	77.20 (1.27)
3hp → 2hp	B	83.12 (1.52)	82.47 (1.72)	83.91 (1.54)	82.83 (0.25)	**85.35** (1.19)	84.92 (1.74)	83.91 (1.52)	83.12 (1.21)	81.53 (0.82)
	OR	77.13 (4.25)	78.43 (3.50)	78.86 (0.45)	77.78 (1.09)	75.61 (2.67)	80.01 (2.53)	80.52 (1.69)	81.24 (0.66)	**81.60** (0.43)

表 3　不同方法在不同未知情景下的安全性平均评估准确率 $\overline{ACC}\uparrow$ 比较

Table 3　The safety average assessment accuracy $\overline{ACC}\uparrow$ comparison of different methods under different unknown scenarios

未知情景	EDA	EADA	RAND	本文方法 (COPOD)	本文方法 (ECOD)	本文方法 (IForest)	本文方法 (KDE)	本文方法 (OCSVM)	本文方法 (LOF)
IR	74.47	74.85	74.51	74.94	75.37	**76.12**	75.36	75.25	75.45
B	83.25	83.02	82.78	83.27	83.55	**83.74**	83.26	82.93	83.11
OR	78.43	78.47	78.10	78.36	77.84	78.67	79.24	**79.36**	78.93

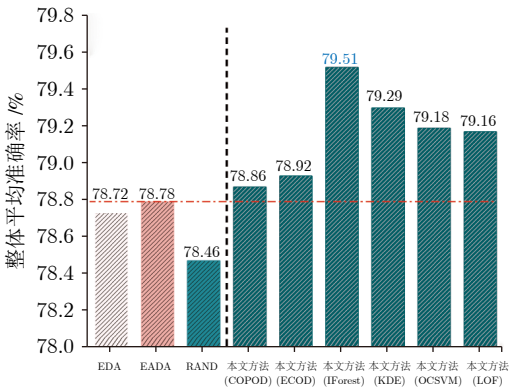

图 5　不同方法的安全性评估整体平均准确率结果对比

Fig. 5　Comparison of overall average accuracy results for safety assessment across different methods

主动域自适应方法都优于传统的域自适应方法. 这说明当目标域中存在未知情景时, 主动标注不一定会带来模型性能的提升, 需要选择合适的样本标注策略. 此外, 现有文献中基于能量模型的主动域自适应方法对模型性能的改善效果并不显著. 相比之下, 本文提出的方法在结合不同分布外检测算法时带来了不同程度的性能提升. 其中, 在与 IForest 算法结合时对模型效果的提升最为显著. 在本文提出的方法中, 分布外检测是关键. 由于不局限于特定的分布外检测算法, 所以本文所提方法具有较大的灵活性和可拓展性. 在未来, 一个值得研究的方向是探索更加有效的分布外检测算法.

4　结束语

本文研究目标域中存在任务相关未知情景的主动域自适应安全性评估问题, 提出一种基于改进能量模型的主动域自适应方法. 通过结合分布外检测器, 该方法能够选择和主动标注目标域中有代表性

的无标注数据, 并加入到训练过程中以提高模型性能. 最后, 基于轴承数据的案例研究说明了所提方法的有效性和可行性. 针对在线安全性评估领域中存在工况变化和任务相关未知情景的挑战, 本文提出的主动域自适应方法提供了一种有前景的解决方案. 未来进一步的研究方向包括开发更有效的分布外检测算法及其在实际动态系统中的应用部署.

References

1 Liu C, He X, Zhou D H, Huang B. Safety assessment for dynamic systems: A survey. *Cybernetics and Intelligence*, DOI: 10.26599/CAI.2024.9390001

2 Wang M, Zhou D H, Chen M Y. Hybrid variable monitoring: An unsupervised process monitoring framework with binary and continuous variables. *Automatica*, 2023, **147**: Article No. 110670

3 Zhang H J, Zhang C, Dong J, Peng K X. A new key performance indicator oriented industrial process monitoring and operating performance assessment method based on improved Hessian locally linear embedding. *International Journal of Systems Science*, 2022, **53**(16): 3538−3555

4 Song P Y, Zhao C H, Huang B. SFNet: A slow feature extraction network for parallel linear and nonlinear dynamic process monitoring. *Neurocomputing*, 2022, **488**: 359−380

5 Liu Qiang, Zhuo Jie, Lang Zi-Qiang, Qin S. Joe. Perspectives on data-driven operation monitoring and self-optimization of industrial processes. *Acta Automatica Sinica*, 2018, **44**(11): 1944−1956
(刘强, 卓洁, 郎自强, 秦泗钊. 数据驱动的工业过程运行监控与自优化研究展望. 自动化学报, 2018, **44**(11): 1944−1956)

6 Zhang Z, He X. Active fault diagnosis for linear systems: Within a signal processing framework. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: Article No. 3505009

7 Amini N, Zhu Q Q. Fault detection and diagnosis with a novel source-aware autoencoder and deep residual neural network. *Neurocomputing*, 2022, **488**: 618−633

8 Xu J M, Ke H B, Chen Z W, Fan X Y, Peng T, Yang C H. Oversmoothing relief graph convolutional network-based fault diagnosis method with application to the rectifier of high-speed trains. *IEEE Transactions on Industrial Informatics*, 2022, **19**(1): 771−779

9 Shakiba F M, Shojaee M, Azizi S M, Zhou M C. Real-time sensing and fault diagnosis for transmission lines. *International Journal of Network Dynamics and Intelligence*, 2022, **1**(1): 36−47

10 Peng Kai-Xiang, Ma Liang, Zhang Kai. Review of quality-related fault detection and diagnosis techniques for complex industrial processes. *Acta Automatica Sinica*, 2017, **43**(3): 349−365
(彭开香, 马亮, 张凯. 复杂工业过程质量相关的故障检测与诊断技术综述. 自动化学报, 2017, **43**(3): 349−365)

11 Gao C, He X, Dong H L, Liu H J, Lyu G R. A survey on fault-tolerant consensus control of multi-agent systems: Trends, methodologies and prospects. *International Journal of Systems Science*, 2022, **53**(13): 2800−2813

12 Jia F L, Cao F F, He X. Adaptive fault-tolerant tracking control for uncertain nonlinear systems with unknown control directions and limited resolution. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, **53**(3): 1813−1825

13 Cai M, He X, Zhou D H. An active fault tolerance framework for uncertain nonlinear high-order fully-actuated systems. *Automatic*, 2023, **152**: Article No. 110969

14 Liu X Q, Chen M Y, Sheng L, Zhou D H. Adaptive fault-tolerant control for nonlinear high-order fully-actuated systems. *Neurocomputing*, 2022, **495**: 75−85

15 Chai Yi, Mao Wan-Biao, Ren Hao, Qu Jian-Feng, Yin Hong-Peng, Yang Zhi-Min, et al. Research on operational safety assessment for spacecraft launch system: Progress and challenges. *Acta Automatica Sinica*, 2019, **45**(10): 1829−1845
(柴毅, 毛万标, 任浩, 屈剑锋, 尹宏鹏, 杨志敏, 等. 航天发射系统运行安全性评估研究进展与挑战. 自动化学报, 2019, **45**(10): 1829−1845)

16 Serbanescu D, Ulmeanu A P. *Selected Topics in Probabilistic Safety Assessment: Methodology and Practice in Nuclear Power Plants*. Switzerland: Springer Nature, 2020. 1−9

17 Aldemir T. A survey of dynamic methodologies for probabilistic safety assessment of nuclear power plants. *Annals of Nuclear Energy*, 2013, **52**: 113−124

18 Zhao Fu-Jun, Zhou Zhi-Jie, Hu Chang-Hua, Chang Lei-Lei, Wang Li. Online safety assessment method based on evidential reasoning for dynamic systems. *Acta Automatica Sinica*, 2017, **43**(11): 1950−1961
(赵福均, 周志杰, 胡昌华, 常雷雷, 王力. 基于证据推理的动态系统安全性在线评估方法. 自动化学报, 2017, **43**(11): 1950−1961)

19 Liu Z Y, Deng Y, Zhang Y, Ding Z J, He X. Safety assessment of dynamic systems: An evidential group interaction-based fusion design. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: Article No. 3523014

20 Liu C, Zhang Y, He X. Expert-augmented data-driven safety level assessment scheme with incremental learning. In: Proceedings of the 12th CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS). Chengdu, China: IEEE, 2021. 1−6

21 Wilson G, Cook D J. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 2020, **11**(5): Article No. 51

22 Zhao Z B, Zhang Q Y, Yu X L, Sun C, Wang S B, Yan R Q, et al. Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study. *IEEE Transactions on Instrumentation and Measurement*, 2021, **70**: Article No. 3525828

23 An Y Y, Zhang K, Chai Y, Liu Q, Huang X H. Domain adaptation network base on contrastive learning for bearings fault diagnosis under variable working conditions. *Expert Systems With Applications*, 2023, **212**: Article No. 118802

24 Chen Z Y, Liao Y X, Li J P, Huang R Y, Xu L, Jin G, et al. A multi-source weighted deep transfer network for open-set fault diagnosis of rotary machinery. *IEEE Transactions on Cybernetics*, 2022, **53**(3): 1982−1993

25 Yang B, Xu S C, Lei Y G, Lee C G, Stewart E, Roberts C. Multi-source transfer learning network to complement knowledge for intelligent diagnosis of machines with unseen faults. *Mechanical Systems and Signal Processing*, 2022, **162**: Article No. 108095

26 Zhu J, Huang C G, Shen C Q, Shen Y J. Cross-domain open-set machinery fault diagnosis based on adversarial network with multiple auxiliary classifiers. *IEEE Transactions on Industrial Informatics*, 2022, **18**(11): 8077−8086

27 Zhang W, Li X, Ma H, Luo Z, Li X. Open-set domain adaptation in machinery fault diagnostics using instance-level weighted adversarial learning. *IEEE Transactions on Industrial Informatics*, 2021, **17**(11): 7445−7455

28 Zhao C, Shen W. Adaptive open set domain generalization network: Learning to diagnose unknown faults under unknown working conditions. *Reliability Engineering & System Safety*, 2022, **226**: Article No. 108672

29 Liu C, Zhang Y, Ding Z J, He X. Active incremental learning for health state assessment of dynamic systems with unknown scenarios. *IEEE Transactions on Industrial Informatics*, 2022, **19**(2): 1863−1873

30 Fu B, Cao Z J, Wang J M, Long M S. Transferable query selection for active domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 7268−7277

31 Prabhu V, Chandrasekaran A, Saenko K, Hoffman J. Active domain adaptation via clustering uncertainty-weighted embeddings. In: Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada:

IEEE, 2021. 8485−8494

32 Xie B H, Yuan L H, Li S, Liu C H, Cheng X J, Wang G R. Active learning for domain adaptation: An energy-based approach. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2022. 8708−8716

33 LeCun Y, Chopra S, Hadsell R, Ranzato M, Huang F J. A tutorial on energy-based learning [Online], available: https://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf, September 8, 2024

34 Liu W T, Wang X Y, Owens J D, Li Y X. Energy-based out-of-distribution detection. arXiv preprint arXiv: 2010.03759, 2021.

35 Li Z, Zhao Y, Botta N, Ionescu C, Hu X Y. COPOD: Copula-based outlier detection. In: Proceedings of the IEEE International Conference on Data Mining. Sorrento, Italy: IEEE, 2020. 1118−1123

36 Li Z, Zhao Y, Hu X Y, Botta N, Ionescu C, Chen G. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022, **35**(12): 12181−12193

37 Liu F T, Ting K M, Zhou Z H. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy: IEEE, 2008. 413−422

38 Desforges M J, Jacob P J, Cooper J E. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 1998, **212**(8): 687−703

39 Scholkopf B, Williamson R C, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1999. 582−588

40 Breunig M M, Kriegel H P, Ng R T, Sander J. LOF: Identifying density-based local outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, USA: ACM, 2000. 93−104

刘 畅 清华大学自动化系博士研究生. 2019 年获得中南大学自动化专业学士学位. 主要研究方向为数据驱动的动态系统安全性评估, 机器学习方法及其工业应用.
E-mail: liuc19@mails.tsinghua.edu.cn
(**LIU Chang** Ph.D. candidate in the Department of Automation, Tsinghua University. He received his bachelor degree in automation from Central South University in 2019. His research interest covers data-driven safety assessment for dynamic systems, machine learning methods, and their applications in industry.)

何 潇 清华大学自动化系长聘教授. 2010 年获得清华大学博士学位. 主要研究方向为动态系统、网络化系统与信息物理系统的故障诊断和容错控制及其应用. 本文通信作者.
E-mail: hexiao@tsinghua.edu.cn
(**HE Xiao** Tenured Professor in the Department of Automation, Tsinghua University. He received his Ph.D. degree from Tsinghua University in 2010. His research interest covers fault diagnosis and fault-tolerant control of dynamic systems, networked systems, cyber-physical systems, and their applications. Corresponding author of this paper.)

王立敏 广州大学机械与电气工程学院教授. 2009 年获得大连理工大学运筹学与控制论专业博士学位. 主要研究方向为批次过程控制, 故障诊断和容错控制.
E-mail: wanglimin0817@163.com
(**WANG Li-Min** Professor at the School of Mechanical and Electrical Engineering, Guangzhou University. She received her Ph.D. degree in operations research and cybernetics from Dalian University of Technology in 2009. Her research interest covers batch process control, fault diagnosis, and fault-tolerant control.)