离线强化学习工业应用

一、赛题背景

随着智能制造浪潮席卷全球，传统工业控制系统在提升生产效率、降低能耗和保障安全生产方面正面临日益严峻的挑战。依赖专家经验与手动调参的传统方法，已难以应对日益复杂多变的工业环境和多种动态干扰。近年来，强化学习在自适应控制领域崭露头角，展现出卓越潜力；但由于在线试验不仅成本高昂，还伴随安全风险，其在实际工业现场的应用受到制约。相比之下，离线强化学习通过充分挖掘历史操作数据与工况记录，在无风险的仿真环境中训练出高效且稳定的控制策略，为工业系统的自动化与智能化升级开辟了全新的路径。

二、赛题应用场景

在钢铁、化工、能源等流程工业中，控制系统必须在产量、能耗与设备寿命等多个控制目标间实现平衡。基于专家经验+PID 的传统调节策略，往往在以下方面力不从心：

1. 复杂耦合下的鲁棒性不足：随着多变量互联、强噪声干扰增多，单纯依赖经验与线性 PID 难以保证系统稳定。

2. 负荷与原料波动中的最优性难以维持：在生产节拍、原料成分频繁变化时，人工整定的控制参数难以快速跟进，导致产量、能效均未达最优。

3. 跨工况迁移的适应性差：每次工艺切换都需耗费大量调试时间，严重制约了产线的灵活调度与快速响应。

离线强化学习（Offline RL）通过深度挖掘既有生产历史数据，并在无风险的仿真或经验回放环境中，自动学习控制策略，具有以下优势：

1. 风险隔离：所有试验均在模拟环境或经验回放池（replay buffer）中进行，无需冒真实设备的安全与停产风险。

2. 高效探索：结合深度网络可以甄别最有价值的历史轨迹，自动挖“少量试错→最大收益”的参数组合。

3. 强迁移能力：学成的策略带有对“状态-动作”映射的普适性认知，能快速应用于不同工况，只需极少量在线微调。

4. 降本增效：避免了反复线下试错与人工整定，实现“即插即用”的智能控制，大幅缩短切换周期、降低能耗并延长设备寿命。

三、出题信息

本赛题由大赛组委会组织专家团队出题，主办单位为南京大学、南栖仙策。主办方出题人：俞扬、秦熔均、高耸屹、屠作霖。

四、赛题任务

本次竞赛将为参赛者提供一个高仿真工业控制模拟器及配套的离线数据集。该数据集通过模拟器与次优控制策略的交互生成，旨在真实还原工业场景中的复杂控制问题。参赛者需要基于提供的离线数据集，设计并训练离线强化学习模型，实现对目标变量的精准控制。控制效果将通过累计奖励值进行评估，最终目标是在以下关键挑战下展现出良好的控制性能与鲁棒性：

1. 时间延迟性：控制指令的效果不会立即显现，存在显著的时间滞后现象；

2. 复杂噪声：观测数据中混合了多种噪声，模拟工业传感器读数中的典型干扰情况；

3. 部分可观测性（POMDP）：智能体仅能获取系统状态的部分信息，需在不完全观测条件下推断全局状态并做出合理决策；

参赛队伍需严格遵循赛方提供的离线数据集进行算法开发，训练策略模型，训练出的控制策略需在模拟环境中实现尽可能高的累计奖励值，以体现其控制性能与鲁棒性。

五、数据集及数据说明

本次竞赛提供的离线数据集由高仿真工业控制模拟器生成。其中初赛数据集仅用于参赛者的算法验证与调试，复赛提供全新数据集（数据格式与初赛保持一致，但任务模拟器的参数发生了变化）。这样是为了能模拟工业界中“模型上线测试机会有限”的挑战，筛选出那些更具实用性和落地能力的算法模型。

数据集详细信息如下：

（一）观测空间

1. 维度：5 维
2. 取值范围：每个观测值均位于[-1, 1]区间
3. 说明：表示系统的各项可观测信息

（二）动作空间

1. 维度：3 维
2. 取值范围：每个动作值均处于[-1, 1]之间

（三）数据结构

1. 轨迹数量：100 条
2. 时间步：每条轨迹包含 1000 个时间步
3. 总数据点：100,000 个（100 轨迹×1,000 时间步）

（四）每个时间步包含的数据

1. obs：当前时间步的观测数据
2. action：次优控制策略所采取的动作
3. next_obs：状态转移后的下一个观测数据

4. reward：当前时间步获得的奖励值

5. index：轨迹标识，同一轨迹内的数据点共享相同的 index

数据集以CSV格式文件提供。示例数据请查看：https://pan.baidu.com/s/1tDFjWFOhAJ8rOBl_HN79ww?pwd=48ux

六、性能指标要求

本次比赛以累计奖励函数值为主要评价指标。参赛者需提交经过训练的智能体策略模型，测试服务器将基于该模型与模拟器进行交互式测试，生成多条运行轨迹。最终，系统将根据所有轨迹的累计奖励值计算平均值，并将其作为模型的最终评分结果。

七、功能要求

算法设计要求：

（一）训练方式：

基于离线数据集完成端到端训练

（二）鲁棒性处理：

1. 具备时间延迟鲁棒性，确保控制决策在不同时间延迟条件下的可靠性
2. 具备噪声鲁棒性，能够有效抑制环境噪声对控制性能的影响
3. 解决状态空间的部分可观测性问题，确保在信息不完全情况下的决策能力

（三）控制性能：

1. 在仿真环境中展现优异的累计奖励值
2. 实现稳定的控制性能，确保系统运行的可靠性

（四）模型特性：

1. 具备良好的泛化能力，确保在不同环境条件下的适应性
2. 具备可扩展性，便于在不同应用场景中部署

八、开发环境

参赛者需使用 Python 进行开发，并将模型上传至模型验证服务平台以获取评分。建议参赛者确保其开发环境与提供测试环境配置一致，以避免因环境差异导致的运行问题。

九、成绩评价

本赛题分为初赛、复赛和决赛三个阶段：

（一）初赛：初赛提供数据集仅供参赛者在前期进行算法验证与调试，选手提交的模型由机器进行自动评分，但不计入最终总分。

（二）复赛：复赛提供新的数据集（格式和初赛保持一致，但任务模拟器的参数会发生变化），选手提交的模型由机器进行自动评分，其得分将计入决赛总分。

（三）线下决赛：线下决赛综合成绩由客观评分和主观评分构成，比例为 70%

和 30%。（国二、国三成绩不涉及 30%主观评分部分）

1. 客观评分：基于经过标准化处理后的复赛机器评测得分；

2. 主观评分：依据经过标准化处理后的答辩得分。答辩评价将综合考察参赛者的答辩表现，以及所提交的技术方案和代码文档。

十、解题思路

（一）策略训练：采用基于离线强化学习算法方法，通过历史数据样本进行策略训练，构建最优控制策略模型。

（二）策略离线评估与调优：可以通过构建高保真的虚拟环境，实现策略的闭环测试与验证。通过实时监控和分析累计奖励指标，建立反馈优化机制，持续改进控制策略。

（三）重点待解决的技术挑战：

1. 针对控制指令时延问题，设计补偿算法；

2. 为应对传感器噪声干扰，进行噪声过滤；

3. 在解决状态空间的部分可观测性问题上，可以引入了基于深度神经网络的状态重构方法。

十一、参考资源

强化学习和离线强化学习相关教材、论文及开源代码库。

十二、提交要求

本次竞赛要求选手提交训练好的策略模型（或推理代码＋模型权重），测试服务器将利用该模型与模拟器进行交互，生成一批轨迹。最终，计算这批轨迹累计奖励的均值作为模型的评分结果。

参赛者请将所有文件打包为一个.zip 文件，顶层目录结构必须如下：

```
submission.zip
├── agent/
│   ├── agent.py          # 策略实现（选手需要继承 BaseAgent 类实现一个 PolicyAgent 类，加载策略模型实现 get_action 函数，文件中已提供了一个模型参考示例供用户参考）
│   ├── model.pth         # 策略模型权重
├── evaluator.py          # 测试模拟器示例，用以帮助测试策略实现是否正确
├── test_agent.py         # 策略测试脚本（不需修改），选手实现 PolicyAgent 类后，可以运行当前脚本和测试模拟器示例进行交互，验证策略实现是否存在问题。
```

上述代码示例仓库会以 git 仓库的形式提供，选手可以下载修改策略实现部分并

打包提交，选手正式报名后即可看到。

资源和格式限制：

1. 单个提交包最大 10MB。若模型权重过大，请自行设计压缩 / 精简方案；
2. 模型测试时间不超过 5 分钟，超时则失败；
3. 模型在离线环境中进行测试，不能访问在线模型；
4. 如果模型比较复杂，建议选手可导出为 ONNX 格式。

十三、更新与答疑

赛题可能会进行更新，为了回复参赛者在参赛过程中遇到的问题，赛题将单独设立选手答疑群（选手正式报名后可看到）。

十四、比赛流程及奖项设置

1. 报名阶段参赛者在比赛官方网站上完成报名注册，提交个人或团队信息，获取初赛数据下载链接。

2. 初赛阶段参赛者利用赛事方提供的训练数据集进行算法模型设计。初赛阶段每个参赛队伍每天仅能提交 1 次。

3. 复赛（省赛）阶段初赛结束后进入复赛阶段，开放复赛数据下载链接。仅有初赛阶段提交有效结果的参赛团队可以进入复赛。复赛期间，参赛者利用赛事方提供的复赛阶段数据进行算法模型调试，提交对复赛测试数据的推理结果。复赛阶段持续 3 天，每个参赛队伍每天仅能提交 1 次。

4. 复赛（省赛）成绩公布在比赛官方网站上公布复赛成绩。以进入复赛参赛团队数量作为计奖基数，按照不超过大赛省赛设奖比例，评选出复赛一、二、三等奖（颁发省赛获奖证书）。评选复赛奖过程中，参赛者提交的算法性能低于赛事方提供的基线参考分数的判定为无效成绩，不予授奖。复赛一、二等奖晋级参加国赛总决赛。

5. 决赛（国赛）阶段

（1）决赛线上评选。晋级决赛的参赛团队，依据复赛排行榜结果，以进入决赛参赛团队数量作为计奖基数，按照不超过大赛国赛设奖比例，评选出国赛一、二等奖候选名单及国赛二、三等奖获奖名单（颁发国赛二、三等奖证书）。

（2）决赛作品提交。国赛一等奖候选团队在规定时间内提交技术文档、算法代码和模型文件、演示视频、补充材料等。提交截止后，不再接受任何形式的修改和补充。

（3）决赛审核阶段。由专业评审团队对国赛一等奖候选参赛团队的参赛作品进行结果复现与审核。评审过程中如有疑问，可要求参赛者进行解释说明。

（4）决赛线下答辩。国赛一等奖候选团队在规定时间内提交完善后的技术文档、算法代码和模型文件、演示视频、补充材料，参加国赛线下总决赛复核答辩，最终依据算法性能得分和线下答辩得分确定国赛一等奖获奖名单及其排名（未参加线下复核

答辩视同放弃奖项）。国赛一等奖颁发荣誉证书。

十五、其它说明

（一）公平性：严禁任何形式的作弊行为，包括但不限于数据泄露、模型预训练数据与测试数据重叠、抄袭他人代码等。一经发现，立即取消参赛资格，并追究相关责任。

（二）知识产权：参赛者提交的作品必须为原创，未在其他比赛中获奖或公开发表。比赛主办方有权对参赛作品进行展示、宣传等相关活动，但知识产权仍归参赛者所有。

十六、联系方式

赛项交流 QQ 群：437743461

邮　　箱：songyi.gao@polixir.ai

报名官网：www.aicomp.cn