2025-06-24

PPO（构思）

- 环境：历史数据
- 状态：启停状态，总流量，总压力
- 策略：PPO ——》选择最佳动作组合
- 奖励：最大化“节能”

状态

- t时刻：
  - 所有机组的启停状态
  - 站房流量
  - 站房压力
- 近几个时刻的总压力，总流量（捕捉趋势）（备选）
- 每台机组的运行时长（备选）
- 归一化

动作

- 机组操作命令：
  - 0：不变
  - 1：改变（开机——》关机，关机——》开机）
- 动作掩码：（约束）
  - 禁止重复启停：（避免对机器的损害）
    - 若刚开机，小于最小开机时间，则不能关机
    - 若刚关机，小于最小关机时间，则不能开机
  - 禁止非法状态：
    - 开机，则不能再开机
    - 关机，则不能再关机

奖励

由于没有直接的单机能耗数据，所以需要设计一个代理奖励。
原则：保留流量和压力前提，尽可能减少运行设备数和启停次数

- 满足流量需求：

    - 流量奖励 = -k1 * max(0,需求流量 - 实际流量)
        - 惩罚流量不满足需求的情况
        - 需求流量基于历史估算
- 满足压力稳定：
    - 压力奖励 = -k2 * (稳定压力 - 实际压力) ^ 2
        - 稳定压力取压力中值
- 减少启停次数：
    - 切换奖励 = -k3 * (启停次数)
        - 来模拟启停成本
- 减少运行台数：
    - 运行设备奖励 = -k4 * (运行的机器数)
- 综合奖励：
    - 综合奖励 = 上述4个求和
- 系数调整：（k1,k2,k3,k4)

环境

基于当前状态，智能体的动作，计算下一刻的状态和奖励

- 下一刻机器状态《——动作掩码《——动作
- 下一刻流量，压力《——历史数据
- 下一刻运行时长，停机时间《——根据状态更新

完整设计流程（简版）

现有的数据：

字段	含义	备注
time	时间	1分钟一个点
DLDZ_AVS_KYJ01_YI01.PV ~ KYJ05_YI01.PV	5台空压机状态（0/1）	每个值：0=停，1=开
DLDZ_DQ200_LLJ01_FI01.PV	站房1瞬时流量	连续值（float）
DLDZ_AVS_LLJ01_FI01.PV	站房2瞬时流量	连续值（float）

模型推荐设计（根据PPO）

项目	含义
模型输入	7维：5个设备状态 + 2个站房流量
模型输出	5维动作：5台空压机的“推荐开关”方案（每台是0或1）
推荐频率	1分钟推荐一次（因原始数据频率为1分钟）
推荐结果	5个值，例如 [1,0,0,1,0]（表示1、4号机开，其余停）

样例

假设：某一分钟数据如下：

字段	当前值
DLDZ_AVS_KYJ01_YI01.PV	1（开）
DLDZ_AVS_KYJ02_YI01.PV	1（开）
DLDZ_AVS_KYJ03_YI01.PV	0（关）
DLDZ_AVS_KYJ04_YI01.PV	0（关）
DLDZ_AVS_KYJ05_YI01.PV	0（关）
DLDZ_DQ200_LLJ01_FI01.PV	150（流量1）
DLDZ_AVS_LLJ01_FI01.PV	130（流量2）

这7个值组成“当前状态”：

```
obs = [1, 1, 0, 0, 0, 150, 130]
```

调用训练好的PPO模型：

```
from stable_baselines3 import PPO
from compressor_env import CompressorEnv

# 加载模型
model = PPO.load("ppo_compressor_model")

# 输入状态（例子）
```

```python
obs = [1, 1, 0, 0, 0, 150, 130]  # 某一分钟真实状态
import numpy as np
obs = np.array(obs).astype(np.float32)

# 预测推荐动作
action, _states = model.predict(obs, deterministic=True)
print("模型推荐的动作：", action)
```

可能模型推荐输出：

```
模型推荐的动作： [1 0 0 1 0]
```

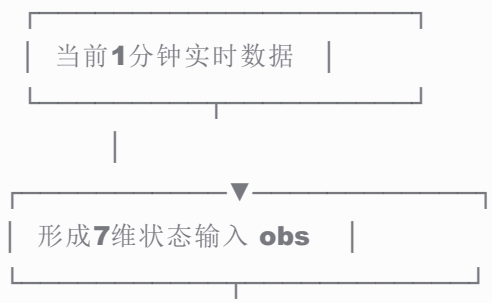
含义：

- 推荐：1号机开，4号机开，其余停。
- 也就是说：
  - 原来1号、2号开机。
  - 现在建议：2号停、4号开。
  PPO背后逻辑：为了满足150+130=280的流量，同时可能发现1号和4号组合更节能或更稳定。

推荐结果的使用流程

步骤	动作	说明
1	读取实时7维状态	1分钟频率
2	**model.predict(obs)**	得到推荐动作
3	比对当前设备状态	如果不一致，建议操作（启/停）
4	执行推荐（如人工审核或PLC自动调整）	实际控制

模型预测完整流程图



训练过程记录

10W步

```
-----------------------------------
| rollout/               |          |
|    ep_len_mean         | 656      |
|    ep_rew_mean         | -2.06e+03 |
| time/                  |          |
|    fps                 | 51       |
|    iterations          | 49       |
|    time_elapsed        | 1948     |
|    total_timesteps     | 100352   |
| train/                 |          |
|    approx_kl           | 0.00564179 |
|    clip_fraction       | 0.0408   |
|    clip_range          | 0.2      |
|    entropy_loss        | -2.65    |
|    explained_variance  | 0.084    |
|    learning_rate       | 0.0003   |
|    loss                | 2.57e+03 |
|    n_updates           | 480      |
|    policy_gradient_loss | -0.00335 |
|    value_loss          | 4.16e+03 |
-----------------------------------
```

用时30分钟+
关键指标：

- ep_rew_mean = -2063 （不能完成任务）
- loss = 2573 （越小越好）
- value_loss =4163 （越小越好）
- explained_variance = 0.084 （远小于1，说明模型能力很差）
- approx_kl = 0.00564179 (太低了， 理想范围：0.01-0.05)

50W步

调参

参数	当前	建议	说明
total_timesteps	100000	500000~1000000	训练步数远远不足，增加
learning_rate	3e-4	1e-4	减小学习率，避免值函数震荡
ent_coef	默认0	0.01~0.05	增加策略熵，提升探索
gamma	0.99	0.95	降低折扣，提升短期奖励影响
gae_lambda	0.95	0.9	提升优势函数稳定性
reward scaling	未做	reward除以10	奖励值尺度过大导致价值函数难以拟合（-2000！）

未运行完，当前状态

```
-------------------------------------
| rollout/                |            |
|    ep_len_mean          | 656        |
|    ep_rew_mean          | -1.96e+03  |
| time/                   |            |
|    fps                  | 52         |
|    iterations           | 120        |
|    time_elapsed         | 4695       |
|    total_timesteps      | 245760     |
| train/                  |            |
|    approx_kl            | 0.0036662796 |
|    clip_fraction        | 0.00347    |
|    clip_range           | 0.2        |
|    entropy_loss         | -3.35      |
|    explained_variance   | 0.887      |
|    learning_rate        | 0.0001     |
|    loss                 | 286        |
```

```
|    n_updates            | 1190        |
|    policy_gradient_loss | -0.00289    |
|    value_loss           | 390         |
-----------------------------------------
```

- total_timesteps = 245760， 当前运行了24W步
- ep_rew_mean = -1963， ↓
- value_loss = 390， 下降很明显了

问题

- [ ] 可能需要重采样（改成1小时）
- [ ] PPO只能根据“输入状态”学习，若输入中没有"变化信号"，模型**根本不知道变化来了**
    - 加入流量变化量：
    - 变化趋势指标：
    - 变化是否超过阈值：
    - 是否处于倒班期