文章编号:1009-1831(2021)04-0073-06　　中图分类号:TM73;TM76　　文献标志码:A

DOI:10.3969/j.issn.1009-1831.2021.04.014

基于深度强化学习算法的“电网脑”及其示范工程应用

徐春雷[1],吴海伟[1],刁瑞盛[2],胡浔惠[3],李　雷[3],史　迪[2]

(1. 国网江苏省电力有限公司,南京　210024;2. 智博能源科技(江苏)有限公司,南京　211302;3. 国电南瑞科技股份有限公司,南京　211106;)

Deep reinforcement learning-based grid mind and field demonstration application

XU Chunlei[1], *WU Haiwei*[1], *DIAO Ruisheng*[2], *HU Xunhui*[3], *LI Lei*[3], *SHI Di*[2]

(1. State Grid Jiangsu Electric Power Co., Ltd., Nanjing 210024, China;2. Zhibo Energy Technology Co., Ltd., Nanjing 211302, China;3. NARI Group Co., Ltd., Nanjing 211106, China)

摘要:可再生能源、电力电子设备渗透率持续增大以及大功率交直流混联,电网的动态性、随机性和不确定性显著增强,给电力系统安全稳定运行带来新的挑战。为更有效解决电网中出现的电压、潮流快速波动而导致的安全问题,提出一种基于最大熵深度强化学习算法的智能电网调控辅助决策方法,同时考虑多种控制目标,对电网运行方式进行在线优化控制。该方法将电网调度控制决策建模为马尔科夫决策过程,训练多线程智能体,并采用周期性在线训练机制对智能体的控制性能进行不断提升。基于该方法所研发的辅助决策原型软件部署在国网江苏电力调度控制中心,可与电网调度控制系统环境直接交互,自主学习且不断提升智能体调控决策能力。训练好的智能体可针对电压越限、联络线潮流越限、网损等综合控制目标在毫秒级时间内给出有效控制策略。

关键词:人工智能;智能调控;深度强化学习;电网安全

Abstract: With the increasing penetration of renewable energy and power electronics-based devices, and the hybrid operation of AC/DC power networks with heavy power transfer, the dynamics, stochastics and uncertainties of the power grid are being observed, threatening its secure operation. In order to effectively resolve security issues caused by fast variations of voltage and line flows, a reinforcement learning algorithm based on maximum entropy depth is presented for providing online decision support in smart grid operation, which can simultaneously consider multiple control objectives. This method formulates decision derivation for grid operation as Markov decision process, which trains multi-threaded soft actor-critic and uses periodic online training mechanism to continuously improve its control performance. The developed prototype using this method has been deployed in the control center of SGCC Jiangsu electric power company, which interacts with live energy management system and learns its control policy adaptively. The well-trained agent can provide effective control actions within milliseconds to regulate voltage violation, line flow and losses.

Key words: artificial intelligence; intelligent dispatch and control; deep reinforcement learning; grid security

0　引言

随着大功率特高压交直流混联,可再生能源渗透率及负荷响应比例逐渐提高,我国电网运行特征发生了深刻且复杂的变化,其不确定性及动态性显著增强。由可再生能源的快速波动以及电网故障所导致的局部功率不平衡,如果没有及时、有效的调控手段,将逐步转变为连锁故障,系统性安全风险显著增大。因此,制定快速、准确的在线调控决策对于确保电网安全稳定运行至关重要。

目前,成功应用于电力领域的人工智能(artificial intelligence, AI)技术多侧重于负荷预测、可再生能源预测、安全性预测等。其核心技术为监督式学习算法,通常需要采集大量有标注的有效样本来训练AI模型。而电网调控领域的很多问题缺少大量真实电网事件作为有效样本,这也是制约监督式学习方法在电网调控领域落地应用的重要因素之一。近期,强化学习算法用于电力领域已有部分研究陆续开展,包括以下方面。

① 电网稳定性控制:文献[1]提出了基于Q学习算法的切机方案来保证系统暂态安全稳定性;文献[2]提出了基于Q学习算法的低频振荡抑制策略。② 微网经济运行:文献[3]提出了在微网环境中基于Q学习算法的储能装置控制方法。③ 提升电网暂态行为指标:文献[4]提出了基于深度Q网络算法的暂态电压控制策略。④ 安全评估:文献[5]提出了使用强化学习算法对电网物理信息系统进行安全评估。⑤ 频率控制:文献[6]提出了使用强化学习进行负荷频率控制的方法。⑥ 电网负荷预测:文献[7]使用强化学习算法进行短期负荷预测。⑦ 经济规划和无功电压控制:文献[8]提出了

收稿日期:2021-03-02;**修回日期**:2021-05-30

基金项目:国网江苏省电力有限公司科技项目(J2020058)

基于分布式强化学习算法来解决动态经济规划的问题；文献[9]提出了一种基于深度强化学习的配电网无功-电压优化方案。⑧ 联络线潮流控制：文献[10]提出了一直基于竞争架构deep Q-learning算法的拓扑控制方法以最大化连续时间断面的线路传输容量；文献[11]提出了一种基于近端优化深度强化学习算法的有功控制方法。⑨ 参数自动调节：文献[12]提出了一种基于多层深度Q网络对发电机动态模型进行自动调参的方法等。

本文在上述研究成果的基础上，提出了一种基于最大熵强化学习算法的电网多目标在线调控辅助决策方法，可对电网有功、无功、网损进行多目标联合优化控制。研发完成的软件部署于江苏省调控中心安全I区，通过多线程离线训练和定期在线更新，训练好的soft actor-critic（SAC）智能体可与电网实时运行环境进行交互，在毫秒级给出辅助调控策略，解决电压越界、联络线潮流越限以及网损优化等问题。该方法利用电力系统基本原理与规则，基于海量电网真实断面进行大量仿真分析，模拟电网中可能出现的电压越界或潮流越限等事件，用于丰富样本库，通过快速自我学习和训练，依靠传统计算分析方法参与评价与反馈，生成满足电网运行控制要求的系列智能体，可对电网中闭环运行的实时调控系统提供有效的辅助支撑，尤其是当闭环调控系统暂时退出运行且调度员缺乏其他有效工具时。

本文首先简述了适用于电网调控领域的深度强化学习基本原理以及本文所使用的最大熵强化学习算法；然后详细给出了所提方法的总体设计、智能体训练流程、原型软件架构以及数据流；最后以江苏张家港分区为例，通过大量的在线数值仿真实验验证了该方法的有效性。

1　深度强化学习技术与最大熵强化学习算法

考虑到实际电网的复杂性，通过对比各算法的优缺点，本文采用最大熵强化学习算法对智能体进行训练以实现既定的控制目标，该算法的鲁棒性和收敛性能十分优异。类似于其他深度强化学习算法（deep reinforcement learning, DRL），SAC也采用值函数和Q函数。区别在于，其他强化学习算法只考虑最大化预期奖励值的积累；而SAC采用随机策略，在最大化奖励值积累的同时最大化熵值，即在满足控制性能要求的前提下采取尽可能随机的控制动作[13]。SAC的核心算法中更新最优策略的过程表示为

$$\pi^{*}=\arg\max_{\pi}\sum_{t}E_{(s_t;\ a_t)\sim\rho_\pi}[R(s_t,a_t)+\alpha H(\pi\cdot|s_t))] \quad (1)$$

式中：$E_{(s_t;\ a_t)\sim\rho_\pi}$为概率分布为$\rho_\pi$采样空间的期望；$R(s_t,a_t)$为$t$时刻状态$s_t$下采取动作$a_t$后得到的奖励值；$H(\pi\cdot|s_t)$为控制策略$\pi$在状态为$s_t$时刻的熵值；$\alpha$为控制探索新控制策略与采用已有控制策略之间的平衡系数。

SAC算法采用随机策略，针对多目标电网自主安全调控这一控制决策问题，具有更强大的探索可行域的能力[13]。训练智能体的过程类似于其他策略梯度算法，对于控制策略的评估和提升可采用带有随机梯度的人工神经网络。构造所需值函数$V_\psi(s_t)$和Q函数$Q_\theta(s_t,a_t)$时，可分别用神经网络参数ψ和θ来表示。SAC算法中采用2个值函数，其中一个值函数称为“软”值函数，来逐步更新策略，以提升算法的稳定性和可靠性。根据文献[13]，软值函数可以通过最小化式(2)中的误差平方值来更新其神经网络的权重，目标函数为

$$J_V(\psi)=E_{s_t\sim D}\left[\frac{1}{2}(V_\psi(s_t)-E_{\alpha_t\sim\pi_\varnothing}[Q_\theta(s_t,a_t)-\alpha\log\pi(a_t|s_t)]\right]^2 \quad (2)$$

式中：D为已有样本的空间分布；$E_{s_t\sim D}$为对误差平方值的期望；$E_{\alpha_t\sim\pi_\varnothing}$为控制策略$\pi_\varnothing$所对应控制动作$a_t$的期望。

式(2)的概率梯度则可用式(3)来计算

$$\hat{\nabla}_\psi J_V(\psi)=\nabla_\psi V_\psi(s_t)\left[V_\psi(s_t)-Q_\theta(s_t,a_t)+\alpha\log\pi_\varnothing(a_t|s_t)\right] \quad (3)$$

式中：∇_ψ为对参数ψ求梯度。

类似地，可通过最小化Bellman残差的方式来更新软Q函数的神经网络权重，计算如下

$$J_Q(\theta)=E_{(s_t,a_t)\sim D}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t)-\hat{Q}(s_t,a_t)\right)\right]^2 \quad (4)$$

$$\hat{Q}(s_t,a_t)=r(s_t,a_t)+\gamma E_{s_{t+1}\sim p}\left[V_{\bar{\psi}}(s_{t+1})\right] \quad (5)$$

式中：γ为折扣系数；$E_{s_{t+1}\sim p}$为概率分布p的$t+1$时刻状态s_{t+1}的期望。

而式(4)的优化求解可由式(6)中的概率梯度进行迭代计算

$$\hat{\nabla}_\theta J_Q(\theta)=\nabla_\theta Q_\theta(s_t,a_t)\left[Q_\theta(s_t,a_t)-r(s_t,a_t)-\gamma V_{\bar{\psi}}(s_{t+1})\right] \quad (6)$$

式中：∇_θ为对θ求梯度；$V_{\bar{\psi}}(s_{t+1})$为目标值函数网络，可定期更新（详见算法1）。

不同于其他确定梯度算法，SAC的策略是由带有平均值和协方差的随机高斯分布所表达。代表其控制策略的神经网络参数可通过最小化预期Kullback-Leibler（KL）偏差而得到，参数为ϕ的控制策略π的目标函数为

$$J_\pi(\phi)=E_{s_t\sim D}\left[E_{a_t\sim\pi_\phi}\left[\alpha\log\left(\pi_\phi(a_t|s_t)\right)-Q_\theta(s_t,a_t)\right]\right] \quad (7)$$

式中：$E_{a_t\sim\pi_\phi}$为控制策略π所对应a_t的期望。

其优化求解过程可由式(8)的概率梯度给出[13]

$$\hat{\nabla}_{\phi}J_{\pi}(\phi)=\nabla_{\phi}\alpha\log\left(\pi_{\phi}(a_t|s_t)\right)+\left(\nabla_{a_t}\alpha\log\left(\pi_{\phi}(a_t|s_t)\right)-\nabla_{a_t}Q(s_t,a_t)\right)\nabla_{\phi}f_{\phi}(\varepsilon_t;s_t) \tag{8}$$

式中：$\hat{\nabla}_{\phi}$为针对$J_{\pi}(\phi)$的ϕ求梯度；∇_{ϕ}为对ϕ求梯度；∇_{a_t}为对a_t求梯度；ε_t为输入向量误差；$f_{\phi}(\varepsilon_t;s_t)$为神经网络变换。训练SAC智能体的算法流程由算法1给出。

2 基于SAC的多目标电网运行方式在线调控方法

2.1 马尔科夫决策过程

电网中的诸多调控问题可描述成马尔科夫决策过程(Markov decision process，MDP)，用于解决随机动态环境下的离散时序控制问题。针对于电网中的电压、潮流控制，相应的MDP过程可用4维元组描述(S，A，P_a，R_a)，其中S代表系统状态空间，可包括电压幅值、电压相角、线路有功功率、线路无功功率、发电机出力、负荷等；A代表控制动作集，可包括发电机有功出力、机端电压设定值、容抗器投切、变压器分接头调整、切负荷等；$P_a(s,s')=Pr(s_{t+1}=s'|s_t=s,\ a_t=a)$则代表系统在$t$时刻从当前状态$s_t$采用了控制动作$a_t$后转移到新状态$s_{t+1}$的概率；$R_a(s,s')$代表从当前状态$s$转移到新状态后$s'$得到的奖励值，用来评估控制效果。

MDP的求解过程是为了得到优化控制策略$\pi(s)$，可从系统状态直接给出控制动作，从而使长时间序列的期望奖励值积累达到最大化。深度强化学习AI智能体可在不断地与环境交互的过程中学习并提升控制策略，即“强化”或“进化”过程，直至快速、高水平完成既定控制目标，如图1所示。通过仔细设计系统状态、奖励值、动作空间，DRL智能体从环境中获取系统状态s，同时给出控制动作a；环境在施加了该控制动作后将改变的系统状态s'和奖励值r输出给智能体。

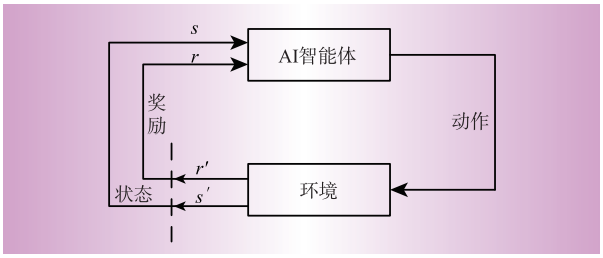

图1 深度强化学习智能体与环境交互过程

Fig. 1 Interaction between DRL agent and environment

在强化学习算法中，有2个重要的函数定义，即值函数和Q函数。其中值函数$V(s)$用来衡量当前状态的好坏，即从当前状态开始并采用一个特定控制策略后所能累计到的奖励值；而Q函数则是用来评估控制策略的好坏，即从某个状态开始采用该控制策略所能积累的奖励值。Q函数为

$$Q^{\pi}(s,a)=E(r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots|s,a) \tag{9}$$

式中：E为对奖励值的期望；r为每个对应时刻或控制迭代所获得的奖励值；γ为折扣系数。

达到最大期望值的最优Q值函数可表述为

$$Q^{*}(s,a)=\max_{\pi}Q^{\pi}(s,a)=Q^{\pi^{*}}(s,a) \tag{10}$$

一旦得到最优Q值函数Q^{*}，AI智能体则可根据该函数给出的值输出控制指令

$$\pi^{*}(s)=\arg\max_{a}Q^{*}(s,a) \tag{11}$$

相应地，最大化奖励值的最优Q值可以表述为

$$Q^{*}(s,a)=r_{t+1}+\gamma\max_{a_{t+1}}r_{t+2}+\gamma^2\max_{a_{t+2}}r_{t+3}+\cdots \tag{12}$$

式(9)至式(12)构成了马尔科夫决策过程。由于控制措施的奖励值可以用人工神经网络来预测，最优的Q值则可以用分解后的形式表述，即贝尔曼(Bellman)方程为

$$Q^{*}(s,a)=E_{s'}[r+\gamma\max Q^{*}(s',a')|s,a] \tag{13}$$

2.2 总体框架

本文提出的方法在训练AI智能体的过程中同时考虑多个控制目标、安全约束和电力设备物理极限。控制目标包括修复电压越限问题、减小网损以及修复联络线潮流越限问题。

值得注意的是，该方法具有通用性和灵活性，可以针对母线电压、联络线功率、线路网损等不同控制问题分别训练、测试AI智能体以提升性能，达到预期的控制目标[14—15]。

2.3 智能体设计

为了训练有效的智能体达到既定目标，相应的环境、样本、状态、动作以及奖励值定义如下。

环境：本文所提出的AI智能体训练方法使用电网真实运行/计算环境，即D5000在线系统中的状态估计模块和调度员交流潮流计算模块。

样本：训练和测试样本可从D5000系统的海量断面潮流文件(QS格式)中获得，代表不同时间点的电网真实运行状态。若针对未来规划中的拓扑结构变化训练AI智能体，则需将该变化反映在样本中。此外，智能体的状态空间和控制空间维度也应进行相应的调整。

状态：针对控制目标，系统状态变量将包括变电站母线电压幅值、电压相角、传输线路有功功率和无功功率、控制变量状态等。

动作：为了有效调整变电站母线电压水平，控制动作可包括调节发电机端电压、投切电容/电抗器、变压器分接头调整、拉停线路等措施。

奖励值：为了施加有效控制，考虑多控制目标后的每一步施加控制措施，所对应的奖励值定义如下。

当发生电压或潮流越限时

$$reward=-\frac{dev\_overflow}{10}-\frac{vio\_voltage}{100} \quad (14)$$

$$dev\_overflow=\sum_{i}^{N}\left(Sline(i)-Sline\_max(i)\right)^2 \quad (15)$$

$$vio\_voltage=\sum_{j}^{M}(V_m(j)-V_{min})\cdot(V_m(j)-V_{max}) \quad (16)$$

式中：N 为功率越限线路的总数；$Sline(i)$ 为线路视在功率；$Sline\_max(i)$ 为线路视在功率极限；M 为电压越限母线的总数；V_m 为母线电压幅值；V_{min} 为电压安全下限；V_{max} 为电压安全上限。

$$delta\_p\_loss=\frac{p\_loss-p\_loss\_pre}{p\_loss\_pre} \quad (17)$$

式中：p_loss 为当前网损值；p_loss_pre 为控制前网损值。

当无电压、潮流越限情况且 $delta\_p\_loss<0$ 时

$$reward=50-delta\_p\_loss\cdot 1\,000 \quad (18)$$

当无电压、潮流越限情况且 $delta\_p\_loss\geqslant 0.02$ 时

$$reward=-100 \quad (19)$$

其他情况时

$$reward=-1-(p\_loss-p\_loss\_pre)\cdot 50 \quad (20)$$

2.4　SAC 智能体训练及测试过程

前期准备工作需要搜集大量代表历史运行工况的电网断面潮流文件，可连续涵盖几周甚至几个月的电网运行状态。

训练开始时，首先提取并解析系统断面潮流文件，由调度员潮流程序进行基态潮流计算并判别是否收敛。若不收敛，则代表该基态潮流文件本身存在数据或模型错误，或电网工况不合理并可能包含安全性问题。若潮流收敛，则分析电网工况，检查包括电压、线路潮流、网损在内的各项指标。提取出的系统状态输入至SAC智能体，给出控制策略。当前样本训练满足退出条件后，将更新SAC的各个神经网络模型参数。当所有样本均被训练后，该流程退出。

为了提高训练效果和控制准确性，通常可以采用多线程训练的方式，即采用不同的超参数和随机数产生多个智能体，综合评估各智能体的效果并选择效果最好的一个或多个，用于在线运行。智能体在测试过程中，SAC智能体的各神经网络模型参数不再改变，而是由训练好的智能体直接给出控制策略，并使用D5000调度员潮流计算程序评估控制效果。

3　江苏电网算例及应用验证

以江苏电网张家港分区为例，分别展示了SAC智能体在2019年夏季高峰典型工况和2019年冬季在线运行的调控性能。

3.1　张家港分区系统简介

图2给出了训练SAC智能体与南瑞D5000系统进行交互的过程。张家港分区的高压网架结构包含45个厂站，线路96条。该分区最大统调出力约230万kW，张家港、晨阳、锦丰主变最大受电能力350万kW，最大供电能力约为580万kW。当D5000系统将断面潮流QS文件输出到AI服务器中，训练好的智能体可在1 s以内给出合理建议来解决电压越界问题并降低系统网损。输出的控制指令将导入D5000系统中进行调度员潮流计算，验证其有效性。图3给出了该原型软件的展示终端界面。

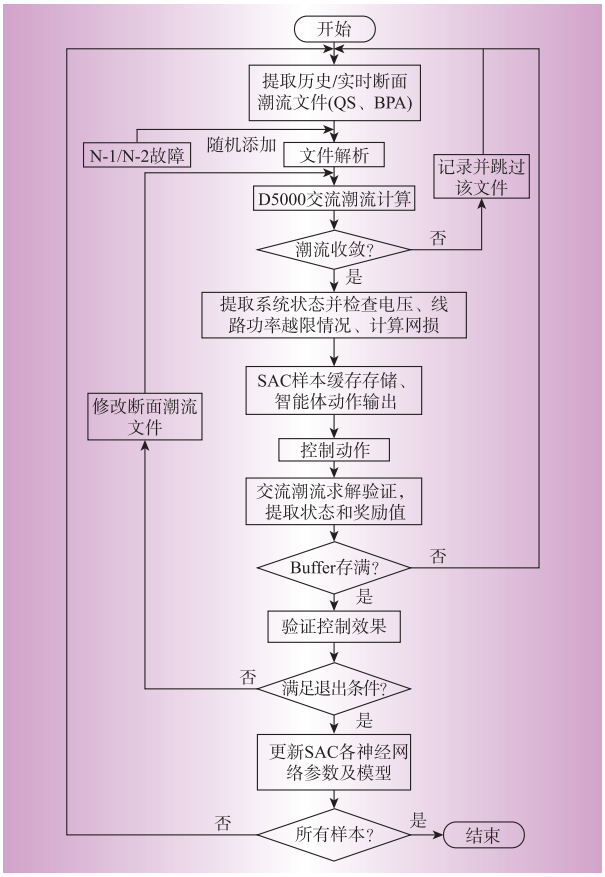

图2　多目标自主调控智能体训练流程图

Fig. 2　Flowchart for training DRL agent for multi-objective autonomous control

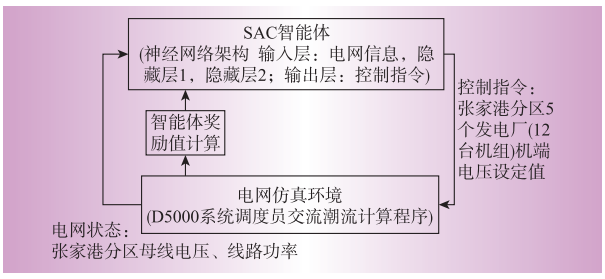

图3　张家港分区AI智能体与电网环境交互过程

Fig. 3　Interaction between SAC agent and power grid environment for Zhangjiagang

该方法在张家港分区的训练与测试分为2个阶段，包括针对典型运行状态的测试和针对在线运行工况的测试。

3.2 典型运行工况测试

在训练该智能体的过程中考虑的控制目标包括：①220 kV及以上母线电压不越限，在[0.97p.u.,1.07p.u.]范围内；② 220 kV及以上线路不过载；③ 降低220 kV及以上线路网损达0.5%以上。控制措施为调节张家港分区内12台发电机的机端电压设定值，在[0.97p.u.,1.07p.u.]范围内调节。训练和测试样本的生成流程如下：在2019年7月份江苏（含全华东地区，220 kV以上网架）5个基态断面潮流文件基础上随机扰动张家港分区负荷（±20%，即80%~120%），并添加N-1、N-1-1故障。共产生了24 000个断面样本，随机选取12 000个作为样本训练SAC智能体，剩余12 000个作为测试样本测试智能体调控性能。

测试结果由表1给出。该测试结果表明经过训练的SAC智能体可以有效帮助典型运行工况缓解电压越限问题及降低网损。结果中存在1个未完全解决电压问题的断面数据，一方面考虑到用于该离线测试的断面数据是在“典型”的实际断面数据上添加各种随机扰动生成的，断面数据本身存在无解的可能性。因此，少量不合理数据本身并不会影响智能体的训练，更重要的是智能体在在线状态下是基于实际数据的测试结果。另一方面，训练和测试智能体过程中遇到难以求解的断面，可以进一步对其进行研究，有可能是电网关键断面。

表1 DRL控制性能总结

Table 1 Summary of DRL control performance

%

控制类别	成功率
电压越限控制	99.991 7
线路过载控制	100.000 0
网损降低	98.330 0

3.3 在线性能测试

本文所研发的软件于2019年11月部署在江苏电网调控中心安全I区。在线系统采用与3.2节相同的控制目标和控制措施。区别在于训练和测试样本均直接从D5000系统的潮流断面QS文件中获得，包括历史断面和实时断面（间隔为5 min）。AI主程序与D5000系统在安全I区实时交互，用来训练和测试智能体的性能。

首先采集2019年11月22日至11月29日的江苏电网断面潮流QS文件对智能体进行训练，其中训练样本1 650个断面数据，测试样本为425个断面数据。智能体的训练和测试性能如图4所示。当施加控制措施后电压和线路功率均不越限，奖励值为正；在此基础上，网损降低越多，奖励值越大。从图5中可以看出，智能体在从零开始训练过程中，前120个断面的效果并不理想，但是随着样本数的增加，其性能不断提升。训练集中共有571个断面出现电压越下限问题，智能体均可以快速且有效地解决；而在测试集中的239个有电压问题的断面均可以有效解决。

相应地，图5给出了智能体训练和测试过程中张家港分区网损降低（输电线路两端有功功率绝对值之差）的情况。在训练集中，智能体可平均降低网损3.453 5%（基准为控制前该分区输电网络网损值）；而在测试集中，智能体可平均降低网损达3.874 7%。

为了确保智能体的控制性能以及避免过拟合情况的发生，每周2次对智能体训练和测试模型进行运维。通过不断积累的训练样本和调试，可保持SAC智能体控制措施的有效性和鲁棒性。表2给出了电网脑系统在2019年12月3日至2020年1月13日期间的运行情况。图6给出了该时间段内张家港分区网损降低情况的总结。

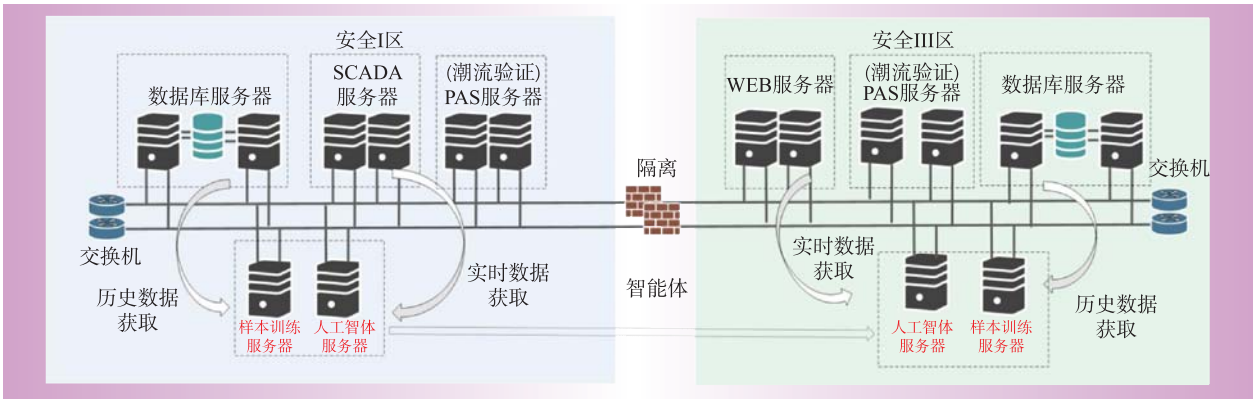

图4 在线系统架构

Fig. 4 Architecture of the online system deployed in Jiangsu province

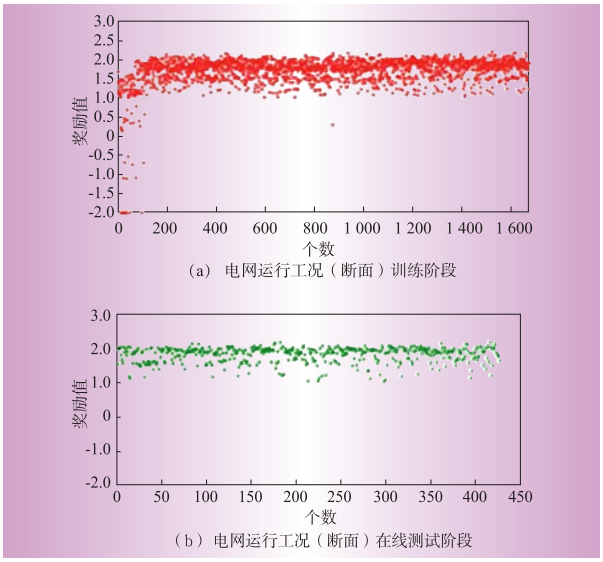

图5　在线系统训练及测试结果

Fig. 5　Performance of training and testing the online system

表2　DRL运行情况总结

Table 2　Summary of DRL operation

内容	结果
有效断面潮流数/个	7 249
网损平均降低/%	3.641 2
线路过载情况/%	0
网损降低/%	98.33
电压越限断面数/个	1 019
电压控制有效率	99.51%完全解决 0.49%有效缓解

本文选取江苏张家港分区进行试运行验证，针对每5 min的电网实时运行断面，SAC智能体在满足调控需求的前提下可在20 ms内对电压、潮流越界等问题提供解决方案，快速消除风险。

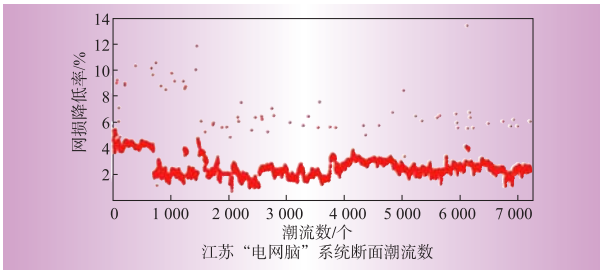

图6　张家港分区网损降低总结

Fig. 6　Summary of network loss reduction in Zhangjiagang

4　结束语

本文介绍了先进人工智能技术在各控制决策领域中的成功应用，阐述了AI技术在电网调控领域的发展瓶颈，讨论了克服该瓶颈的方法和思路，并在此基础上提出基于深度强化学习算法的多目标多工况电网在线优化控制方法。本文所述方法是人工智能DRL技术在实际电力系统调控领域的应用实践。测试结果和试运行性能说明，基于人工智能技术的电力系统控制和优化具有广阔前景。

参考文献：

[1] 刘威，张东霞，王新迎，等. 基于深度强化学习的电网紧急控制策略研究[J]. 中国电机工程学报，2018，38(1)：109-119.
LIU Wei，ZHANG Dongxia，WANG Xinying，et al. A decision making strategy for generating unit tripping under emergency circumstances based on deep reinforcement learning [J]. Proceedings of the CSEE，2018，38(1)：109-119.

[2] DUAN J，XU H，LIU W. Q-learning based damping control of wide-area power systems under cyber uncertainties [J]. IEEE Transactions on Smart Grid，2018，9(2)：6 408-6 418.

[3] DUAN J，YI Z，SHI D，et al. Reinforcement-learning-based optimal control for hybrid energy storage systems in hybrid AC/DC microgrids [J]. IEEE Transactions on Industrial Informatics，2019，15(9)：5 355-5 364.

[4] HUANG Q，HUANG R，HAO W，et. al. Adaptive power system emergency control using deep reinforcement learning[J]. IEEE Transactions on Smart Grid，2019，11(2)：1 171-1 182.

[5] LIU C，KONSTANTINOU C. Reinforcement learning for cyber-physical security assessment of power systems [C]. IEEE Milan Power Tech Conference，June 23-27，2019，Milano，Italy.

[6] YAN Z，XU Y. Data-driven load frequency control for stochastic power systems：a deep reinforcement learning method with continuous action search [J]. IEEE Transactions on Power Systems，2019，34(2)：1 653-1 656.

[7] FENG C，ZHANG J. Reinforcement learning based dynamic model selection for short-term load forecasting [C]. IEEE PES Innovative Smart Grid Technologies Conference (ISGT)，February 18-21，2019，Washington DC，USA.

[8] DAI P，YU W，WEN G，et al. Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions [J]. IEEE Transactions on Industrial Informatics，2020，16(4)：2 258-2 267.

[9] WANG W，YU N，GAO Y，et al. Safe off-policy deep reinforcement learning algorithm for volt-VAR control in power distribution systems [J]. IEEE Transactions on Smart Grid，2020，11(4)：3 008-3 018.

[10] LAN T，DUAN J，Zhang B，et al. AI-based autonomous line flow control via topology adjustment for maximizing time-series ATCs [C]. IEEE PES General Meeting，August 3-6，2020，online.

[11] ZHANG B，LU X，DIAO R，et al. Real-time autonomous line flow control using proximal policy optimization [C]. IEEE PES General Meeting，August 3-6，2020，online.

[12] WANG S，DIAO R，LAN T，et al. A DRL-aided multi-layer stability model calibration platform considering multiple events [C]. IEEE PES General Meeting，August 3-6，2020，online.

[13] HAARNOJA T，ZHOU A，ABBEEL P，et al. Soft actor-critic：off policy maximum entropy deep reinforcement learning with a stochastic actor [C]. Proceedings of Machine Learning Research，July 2-5，2018，Stockholm Sweden：1 861-1 870.

[14] DIAO R，WANG Z，SHI D，et al. Autonomous voltage control for grid operation using deep reinforcement learning [C]. IEEE PES General Meeting，August 4-8，2019，Atlanta，USA.

[15] DUAN J，SHI D，DIAO R，et al. Deep-reinforcement-learning-based autonomous voltage control for power grid operations [J]. IEEE Transactions on Power Systems，2019，35(1)：814-817.

（责任编辑　郝　洁）