文章编号:1674-0629(2024)11-0067-12 中图分类号:TM711;TM727 文献标志码:A
DOI:10. 13648/j. cnki. issn1674-0629. 2024. 11. 008

基于图神经网络与强化学习的配电网电压与无功功率优化方法

朱涛[1],海迪[2],李文云[3],黄伟[3],周胜超[2],吴明贺[4],王逸飞[4]
(1. 云南电网有限责任公司红河供电局,云南 红河 661100;2. 昆明供电局电力调度控制中心,昆明 650011;3. 云南电网有限责任公司,昆明 650217;4. 东南大学电气工程学院,南京 210096)

摘要:高比例分布式光伏的接入改变了配电网的运行方式,并导致配电网出现有功功率损耗过大、调压设备寿命下降、节点电压越限等一系列问题。基于此背景,首先将电压无功优化问题建模为一个马尔科夫决策过程,然后使用无模型的深度强化学习方法进行求解,该方法可以从历史运行数据中捕获光伏发电的间歇性和负荷的波动性。提出了一种图卷积网络的近端策略优化算法(graph convolutional network-proximal policy optimization,GCN-PPO),该算法通过嵌入图卷积网络来提高强化学习智能体对配电网图数据的感知能力。最后以改进的IEEE 33节点测试系统开展算例分析,验证了所提方法的有效性和相比其他方法的优势。结果表明,基于图卷积网络训练的强化学习智能体在配电网拓扑发生变化和测量数据丢失时仍表现出较好的性能。

关键词:分布式光伏;配电网;电压无功优化;深度强化学习;图卷积网络

Voltage and Reactive Power Optimization Method for Distribution Networks Based on Graph Neural Network and Reinforcement Learning

ZHU Tao[1], HAI Di[2], LI Wenyun[3], HUANG Wei[3], ZHOU Shengchao[2], WU Minghe[4], WANG Yifei[4]
(1. Honghe Power Supply Bureau, Yunnan Power Grid Co., Ltd., Honghe, Yunnan 661100, China; 2. Power Dispatching and Control Center, Kunming Power Supply Bureau, Kunming 650011, China; 3. Yunnan Power Grid Co., Ltd., Kunming 650217, China; 4. School of Electrical Engineering, Southeast University, Nanjing 210096, China)

Abstract: High propotional distributed photovoltaic integration changes the operation mode of the distribution networks, and leads to a series of problems such as excessive active power losses, reduced service life of regulating equipment, and exceeding node voltage limits in the distribution networks. Based on this background, firstly the voltage and reactive power optimization problem is modelled as a Markov decision process, which is solved by using a model-free deep reinforcement learning method that captures the intermittence of PV and load fluctuation from historical operating data. A graph convolutional network-proximal policy optimization (GCN-PPO) algorithm is proposed which improves the perception of reinforcement learning agent on graph data of distribution networks by embedding the graph convolutional network. Finally, an arithmetic analysis is carried out with a modified IEEE 33-node test system to verify the effectiveness of the proposed method and its advantages over other methods. The results show that the trained reinforcement learning agent based on graph convolutional networks exhibits better performance when the topology of the distribution network changes and the measurement data are lost.

Key words: distributed photovoltaics; distribution networks; voltage and reactive power optimization; deep reinforcement learning; graph convolutional networks

基金项目:国家自然科学基金资助项目(52007032)。
Foundation item: Supported by the National Natural Science Foundation of China (52007032).

0 引言

在“碳达峰，碳中和”的背景下，我国分布式发电的规模不断扩大。据能源局统计，截止2021年底分布式光伏的装机容量达到10.75 GW，约占全部并网光伏容量的三分之一[1]。高比例光伏(photovoltaics，PV)的渗透会给配电网的稳定运行带来一定的挑战，具体表现在有功损耗增加、节点电压越限、离散调压设备抽头频繁变动导致的寿命下降等等。同时由于依赖天气的PV发电具有间歇性和波动性，也给问题的求解增加了复杂性和不确定性[2]。

电压无功优化(Volt/Var optimization，VVO)任务可以被描述为一个混合整数非线性规划问题，其变量一般代表电压调节设备的离散动作和无功功率补偿设备的连续动作。目前的研究热衷于通过预测的方式或者使用随机概率函数来表征PV的不确定性。比如，文献[3]使用模型预测控制方法来求解高比例的分布式电源高渗透时的VVO问题。文献[4]开发了一种两阶段的随机规划模型，然后再将该模型转换成一个确定性的混合整数二次规划模型进行求解。与随机规划不同，鲁棒优化方法通过构造不确定性集的方式来表征PV出力的随机性[5]。文献[6]使用鲁棒优化模型制定了VVO任务，然后通过一种乘法器交替方向法进行求解。文献[7]针对慢速离散调压设备制定了一种日前-日内两级无功功率滚动调度方法，并在日前尺度上建立了一种两阶段鲁棒优化模型。然而鲁棒优化的主要思想是在最坏的情况下进行求解，其结果通常是保守的。除此之外，像粒子群优化(particle swarm optimization，PSO)算法[8]和遗传算法[9]等启发式方法也常被用来求解配电网的VVO问题。文献[10]通过调度静止无功功率补偿器(static var compensators，SVC)来求解含分布式电源配电网的VVO问题，并提出了一种改进的粒子群优化(particle swarm optimization，PSO)算法进行全局寻优。文献[11]提出了一种考虑电压越限风险的长时间尺度VVO模型，并使用灾变遗传算法对模型进行求解。

虽然传统的数学优化方法可以求得高质量的解，但是其计算负担过大，最佳指令不能在规定的时间内下发给设备。启发式方法在一定程度上减小了计算量，但所求的解不能保证是全局最优的。同时，上述方法在处理新能源不确定性方面是极不准确的[12]。为了克服这些挑战，基于深度强化学习(deep reinforcement learning，DRL)的无模型的方法是一种良好的替代方案。强化学习智能体可以在合适的奖励函数下，通过不断与环境互动学习系统的动态特性从而提供高质量的决策方案。与基于模型的方法不同，DRL方法可以在历史运行数据中捕获系统的不确定性和随机性，训练出的智能体具有一定的泛化能力[13]。同时，经过精心的训练和设计，DRL算法可以实时做出决策以避免在线计算时间过长。

目前许多研究都使用了DRL方法来解决VVO问题，并取得了不错的效果。文献[14-16]使用了单智能体的DRL模型，将配电网的无功优化问题建模为一个马尔科夫决策过程(Markovian decision process，MDP)。文献[14]采用行动者-评论家(actor-critic，AC)的算法把最小化网损和设备的动作成本作为优化目标，以离散无功功率调节设备的投切指令为控制变量进行VVO的求解。然而为了应对PV的间歇性带抽头的离散设备要经常动作，这导致了其使用寿命的降低。为了解决这个问题，文献[15]采用了SVC和PV附带的逆变器对配电网进行无功功率补偿，并使用软行动者-评论家(soft actor-critic，SAC)算法为智能体指定独特的价值和策略模型。文献[16]使用近端策略优化(proximal policy optimization，PPO)算法来求解具有可再生能源发电和储能设备的不确定性最优潮流问题。文献[17]提出了一种多时间尺度的VVO问题，并使用多智能体深度确定性策略梯度算法进行求解。文献[18-19]首先将配电网划分为多个子网络，将每个子网络看作一个智能体，然后把多区域的VVO问题描述为一个马尔科夫博弈模型并使用多智能体强化学习算法进行求解。

在上述DRL文献中，电力系统中的电气量以向量的形式输入到神经网络来训练智能体。然而电力系统中的稳态数据是包含节点、边的电气量和拓扑连接关系的图数据结构，属于非欧数据。图神经网络(graph neural networks，GNN)是一类直接在图形数据结构上操作的神经网络，其核心思想是利用消息传递机制来聚合相邻节点的特征。最近的研究表明，将图神经网络作为强化学习的感知层可以提高智能体的训练稳定性和在不同拓扑结构上的扩展

能力。文献[20]提出一种基于稳态数据的电力系统暂态稳定评估方法，使用GNN对电网的图数据和拓扑连接关系进行建模。文献[21]提出了一个电网故障恢复的图强化学习框架，使用图卷积网络(graph convolutional network，GCN)捕捉了电力网络中故障恢复的复杂机制。文献[22]使用基于GNN的DRL方法来解决配电网的电压无功控制问题，并证明了基于图的策略模型对图中错误拓扑信息的鲁棒性更加稳健，然而该研究并没有考虑PV接入的场景，且所提出的图强化学习方法并不适用于构建PV和SVC智能体。本文提出了一个将DRL和GCN相结合的框架来解决PV高渗透配电网快时间尺度的电压控制问题，并证明了基于图神经网络训练的DRL模型具有较强的可扩展性。

本文的创新点和贡献如下：1)将传统的VVO问题建模为一个MDP，并使用基于近端策略优化的DRL方法来训练具有连续无功功率补偿能力的智能体，使其可以快速应对高比例PV和负荷的间歇性以及波动性；2)提出了一种GCN-PPO算法，将GCN架构融入到PPO算法的策略(actor)和价值(critic)模型之中，提高了模型对电力系统图数据的感知能力。3)与传统方法在改进的IEEE-33节点测试系统上进行对比，验证了所提方法的VVO性能、鲁棒性和可扩展能力。

1　配电网电压无功优化模型

配电网在运行时通常呈现径向的网络拓扑，本文使用图G=(N，E)来表示一个配电网，N和E分别为配电网所有节点和线路的集合，对于有节点集合为N的配电网来说，节点$i\in N$。本文考虑的不确定量包括PV的有功出力P_{PV}、负荷的有功功率P_{L}和负荷的无功功率Q_{L}，在执行最优潮流(optimal power flow，OPF)时的控制变量为SVC的无功功率输出Q_{SVC}和PV逆变器的无功功率输出Q_{PV}。下面将阐述VVO的优化目标和约束条件。

本文中无功优化的目标是最小化配电网的有功功率损耗，式(1)为t时刻的优化目标，其网损的计算如式(2)所示[16]。

$$\min_{Q_{\mathrm{SVC}},Q_{\mathrm{PV}}} F=\min\sum_{t=1}^{T}C_{\mathrm{p}}P_{\mathrm{loss}}(t) \tag{1}$$

$$P_{\mathrm{loss}}(t)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}G_{ij}\left[V_{\mathrm{e},i}^{2}(t)+V_{\mathrm{f},i}^{2}(t)+V_{\mathrm{e},j}^{2}(t)+V_{\mathrm{f},j}^{2}(t)-2(V_{\mathrm{e},i}(t)V_{\mathrm{e},j}(t)+V_{\mathrm{f},i}(t)V_{\mathrm{f},j}(t))\right] \tag{2}$$

式中：F为在时间段T内的总网损成本；$P_{\mathrm{loss}}(t)$为配电网t时刻的网络损耗；ij为第ij节点间的线路；C_{p}为网损成本系数；G_{ij}为第ij节点间线路的导纳矩阵的实部；$V_{\mathrm{e},i}(t)$和$V_{\mathrm{f},i}(t)$分别为t时刻内i节点电压的实部和虚部；$V_{\mathrm{e},j}(t)$和$V_{\mathrm{f},j}(t)$分别为t时刻内j节点电压的实部和虚部；N为配电网节点的集合。

本文VVO的约束条件大致可以分为3类：潮流计算的等式约束和不等式约束以及调压设备运行容量约束，其表达式如下。

$$P_i(t)+V_{\mathrm{e},i}(t)\sum_{j=1}^{N}(G_{ij}V_{\mathrm{e},j}(t)-B_{ij}V_{\mathrm{f},j}(t))+V_{\mathrm{f},i}(t)\sum_{j=1}^{N}(G_{ij}V_{\mathrm{f},j}(t)-B_{ij}V_{\mathrm{e},j}(t))=0,\quad i\in N \tag{3}$$

$$P_i(t)=P_{\mathrm{L},i}(t)-P_{\mathrm{PV},i}(t),\qquad i\in N \tag{4}$$

$$Q_i(t)+V_{\mathrm{f},i}(t)\sum_{j=1}^{N}(G_{ij}V_{\mathrm{e},j}(t)-B_{ij}V_{\mathrm{f},j}(t))-V_{\mathrm{e},i}(t)\sum_{j=1}^{N}(G_{ij}V_{\mathrm{f},j}(t)+B_{ij}V_{\mathrm{e},j}(t))=0,\quad i\in N \tag{5}$$

$$Q_i(t)=Q_{\mathrm{L},i}(t)-Q_{\mathrm{SVC},i}(t)-Q_{\mathrm{PV},i}(t),\qquad i\in N \tag{6}$$

$$V_{\min}<V_i(t)<V_{\max},\qquad i\in N \tag{7}$$

式中：B_{ij}为第ij节点间线路导纳矩阵的虚部；$V_{\mathrm{e},i}(t)$和$V_{\mathrm{f},i}(t)$分别为t时刻内i节点电压的实部和虚部；$V_{\mathrm{e},j}(t)$和$V_{\mathrm{f},j}(t)$分别为t时刻内j节点电压的实部和虚部；$P_i(t)$和$Q_i(t)$分别为t时刻节点i有功功率和无功功率的净注入量；$P_{\mathrm{L},i}(t)$和$Q_{\mathrm{L},i}(t)$分别为t时刻节点i负荷的有功功率和无功功率；$P_{\mathrm{PV},i}(t)$和$Q_{\mathrm{PV},i}(t)$分别为t时刻节点i光伏输出的有功功率和无功功率；$Q_{\mathrm{SVC},i}(t)$为t时刻节点i SVC设备输出的无功功率；式(3)—(6)为有功和无功的功率平衡方程；式(7)为节点电压的上下限约束，其中$V_i(t)$为t时刻节点的电压幅值；$V_{\max}$和$V_{\min}$分别为节点电压上下限，一般设置为1.07 p.u.和0.97 p.u.。

$$Q_{\mathrm{SVC},\min}\leqslant Q_{\mathrm{SVC},i}(t)\leqslant Q_{\mathrm{SVC},\max} \tag{8}$$

$$Q_{\mathrm{PV},\min}(t)\leqslant Q_{\mathrm{PV},i}(t)\leqslant Q_{\mathrm{PV},\max}(t) \tag{9}$$

式中：$Q_{\mathrm{SVC},\min}$和$Q_{\mathrm{SVC},\max}$分别为SVC的容量上下限；$Q_{\mathrm{PV},\min}(t)$和$Q_{\mathrm{PV},\max}(t)$分别为t时刻PV无功功率输出的上下限。考虑到安全性，PV逆变器通常设有冗余的额定容量，并在最大功率点跟踪模式下运行；

式(8)—(9)分别为SVC和PV逆变器进行无功功率调节时的调节范围。

因此逆变器的无功功率控制范围可以由PV的额定装机容量S_{PV}和当前的有功输出$P_{\mathrm{PV}}(t)$决定，$S_{\mathrm{PV},i}$代表第i个节点上PV的容量大小，t时刻节点i上PV输出的无功功率输出范围为：

$$-\sqrt{S_{\mathrm{PV},i}^2-P_{\mathrm{PV},i}^2(t)}\leqslant Q_{\mathrm{PV},i}(t)\leqslant\sqrt{S_{\mathrm{PV},i}^2-P_{\mathrm{PV},i}^2(t)} \tag{10}$$

2 基于GCN-PPO算法的电压无功优化方法

2.1 强化学习与图神经网络

强化学习是一种数据驱动的方法，它解决了控制动态系统的问题。强化学习智能体通过不断地和仿真环境进行互动来捕获所求问题的动态特性，然后根据历史经验在奖励值的指导下学习相应的策略。整个过程通常被建模为一个MDP，本论文使用一个五元组$\mathcal{M}(\mathcal{S},\mathcal{A},\mathcal{P},r,\gamma)$来表示一个MDP，其中$\mathcal{S}$是状态$s$的集合($s\in$S)，$\mathcal{A}$是动作$a$的集合($a\in\mathcal{A}$)。$\mathcal{P}$是状态转移概率，它通过一个概率分布$P\left(s_{t+1}|s_t,a_t\right)$来描述系统的动态，$s_t$为系统$t$时刻的状态，$a_t$为系统$t$时刻的动作，$s_{t+1}$为系统$t$+1时刻的状态。$r$是奖励函数($\mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$)，它用来指导智能体学习正确的策略。$\gamma\in(0,1]$是折扣系数，它将奖励和时域相关联使智能体考虑了未来的回报。

强化学习的最终目标是学习一个策略$\pi(a_t\mid s_t)$，该策略就是某一状态所对应的动作分布。本文使用一个轨迹列表κ来表示长度为N的状态和动作序列，其中$\kappa=(s_0,a_0,\cdots,s_N,a_N)$。对于一个给定的M和策略$\pi$，其轨迹分布为：$P_\pi(\kappa)=\Pi_{t=o}^{T}\pi\left(a_t|s_t\right)P\left(s_{t+1}|s_t,a_t\right)$。强化学习的目标$J(\pi)$可以被描述为上述轨迹分布下的期望值，$T$为智能体控制视界的长度，即$J(\pi)=\mathbb{E}_{\kappa\sim P_\pi(\kappa)}\left[\sum_{t=0}^{T}\gamma^t r\left(s_t,a_t\right)\right]$，该目标提供了一个更直观的数学形式，即通过最大化轨迹分布下的预期累计奖励之和来实现基于强化学习的控制[23]。通常情况，强化学习使用策略梯度理论来优化目标$J(\pi)$，同时使用θ来参数化策略$\pi_\theta(a_t\mid s_t)$。因此，可以使用式(11)来表示策略梯度。

$$\nabla_\theta J(\theta)=\mathbb{E}_{\kappa\sim p_{\pi_\theta}(\kappa)}\left[\sum_{t=0}^{T}\gamma^t\nabla_\theta\log\pi_\theta(a_t|s_t)Q\left(s_t,a_t\right)\right] \tag{11}$$

上式中的状态动作值函数$Q(s_t,a_t)$也可以使用一个神经网络进行参数化，其计算式如下：

$$Q\left(s_t,a_t\right)=\sum_{t'=t}^{N}\gamma^{t'-t}r(s_{t'},a_{t'})-V(s_t) \tag{12}$$

式中$V(s_t)$为状态s_t下的状态值函数，通常被表示为[24]：

$$V(s_t)=\mathbb{E}_{\kappa\sim p_\pi(\kappa|s_t)}\left[\sum_{t'=t}^{N}\gamma^{t'-t}r(s_{t'},a_{t'})\right] \tag{13}$$

图神经网络是一类特殊的神经网络，它可以从图形结构的数据中捕捉不同节点之间的关系。GNNs的核心思想是利用消息传递机制来聚合相邻节点的特征。目前大多数GNNs模型可以被看作是一个学习“置换不变性函数”的方法，它的输入是一个特征矩阵$\boldsymbol{X}$和一个连接矩阵$\boldsymbol{A}$。$\boldsymbol{X}$描述了每个节点的特征$\boldsymbol{x}_i$，其维度为节点的特征数乘以节点数，同时用$\boldsymbol{x}'_i$表示经过神经网络更新后的下一层特征，$\boldsymbol{A}$表示了节点之间的连接关系。本文使用文献[25]提出的GCN来表征G的信息，其核心是使用一个图卷积算子描述的函数$f(\boldsymbol{X},\boldsymbol{A})$来进行有效的信息传递。式(14)描述了GCN的信息传递规则。

$$\boldsymbol{X}'=f(\boldsymbol{X},\boldsymbol{A})=\sigma\left(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}\right) \tag{14}$$

式中：$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$($\boldsymbol{I}$为单位矩阵)，$\hat{\boldsymbol{D}}$为$\hat{\boldsymbol{A}}$的对角度矩阵；$\sigma(\cdot)$为激活函数；$\boldsymbol{W}$为全连接层的参数矩阵。

总结来说，GCN描述了一个置换不变的传播规则，该规则通过聚合相邻节点的信息来更新节点的特征。

2.2 电压无功优化的马尔科夫决策过程

本节将配电网的VVO模型描述为一个MDP模型，其中状态空间、动作空间、奖励函数和状态转移过程如下。

1) 状态空间

本文将向量$\boldsymbol{x}_i=(P_i(t),Q_i(t),V_i(t))$作为$t$时刻节点$i$的特征值，分别为节点$i$的有功功率和无功功率的净注入量以及电压幅值，并假设每个节点的特征都可以被观测到。在本文中，特征矩阵和连接矩阵被作为状态量输入图卷积神经网络中。t时刻的状态空间可以表示为：$s_t=(\boldsymbol{X}_t,\boldsymbol{A}_t)$，$\boldsymbol{X}_t$为$t$时刻所有节点特征组成的特征矩阵($\boldsymbol{X}$的维度为3×$N$)，$\boldsymbol{A}_t$为$t$时刻配电系统所有线路的连接关系。

2) 动作空间

智能体的的动作包括了配电网中所有无功功率

调节设备的无功功率输出，假设有 m 个SVC和 n 个PV逆变器，则 $a_t=\{Q_{\mathrm{SVC},i}(t),Q_{\mathrm{PV},j}(t)\}$，$i\in n$，$j\in m$。

3）奖励函数

与文献[13]相同，本文将电压越限作为惩罚项加入奖励函数来使智能体在学习的过程中执行电压安全约束，t 时刻的奖励函数为：$r_t=C_{\mathrm{p}}P_{\mathrm{loss}}(t)+C\mathrm{v}$，其中 $C\mathrm{v}$ 为电压越限惩罚系数。

4）状态转移过程

在每个时间步 t 内，智能体观测当前的状态 s_t，然后根据 s_t 做出当下的动作 a_t，最后获得一个奖励值 r_t 并根据P得到下一时刻的状态 s_{t+1}。智能体的目标就是通过以上过程寻找一个最大化累计期望回报 $\sum^{N}_{t=0}\gamma^t r(s_t,a_t)$ 的策略。

2.3　基于GCN的近端策略优化算法

目前强化学习方法可以被分为三类：基于价值（value-based）的方法、基于策略（policy-based）的方法和基于AC框架的算法。本论文使用了基于AC框架设计的近端策略优化（PPO）算法[26]，它具有随机策略稳定性高的优点。

本论文提出的GCN-PPO算法在多层感知机（multi-layer perception，MLP）神经网络前面增加了图卷积层，提高了PPO智能体对图数据的感知能力。图1展示了GCN-PPO算法中Actor网络和Critic网络的结构。本文使用的Actor网络架构由两个图卷积层和3个MLP层组成，每一层均附带一个ReLU激活函数，并使用求和池化函数将图卷积层的输出在相邻的节点上聚合起来，然后传递给MLP层输出一个策略。定义值函数的Critic网络的架构和Actor网络大致相同，主要的区别是在其图卷积层后面加了一个全局求和池函数，这样使得价值函数能够聚合图中所有节点的信息，从而计算整个网络的估计值。

PPO算法是在信任域策略优化（trust region policy optimization，TRPO）的基础上改进得来的。TRPO算法使用库尔贝克-莱布朗（Kullback-Leibler，KL）散度约束策略网络使其更新的策略接近于旧策略，其优化目标和约束如式（15）—（16）所示。

$$\max_{\theta}\hat{\mathbb{E}}_{\pi_{\theta_{\mathrm{old}}},t}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}\hat{A}_{\pi_{\theta_{\mathrm{old}}},t}\right]=\hat{\mathbb{E}}_{\pi_{\theta_{\mathrm{old}}},t}\left[r_t(\theta)\hat{A}_{\pi_{\theta_{\mathrm{old}}},t}\right]\quad(15)$$

$$\text{s.t. }\hat{\mathbb{E}}_{\pi_{\theta_{\mathrm{old}}},t}\left[\mathrm{KL}\left[\pi_{\theta_{\mathrm{old}}}(\cdot|s_t),\pi_\theta(\cdot|s_t)\right]\right]\leqslant\delta\quad(16)$$

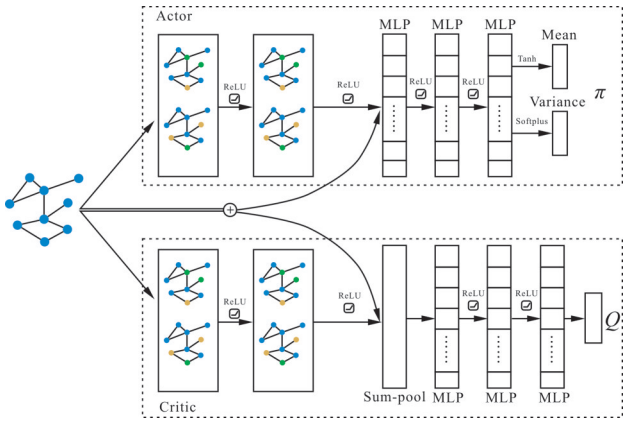

图1　GCN-PPO的Actor和Critic网络结构

Fig. 1　Actor and Critic network structure of GCN-PPO

式中：$r_t(\theta)=\dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$ 为新旧策略的比率；$\pi_{\theta_{\mathrm{old}}}$ 为更新前的旧策略；θ 为策略参数；KL散度也可以被称作相对熵，用来衡量概率分布之间的差异；δ 为置信度，用于限制策略的更新幅度；$\hat{\mathbb{E}}_{\pi_\theta,t}[\cdot]$ 为期望，表示在有限样本上的经验平均；$\hat{A}_{\pi_{\theta_t}t}$ 为在策略 π_θ 下 t 决策步的优势函数估计值。

由于在每次策略更新中计算KL散度的计算成本非常高，因此PPO算法采用了截断函数代替KL散度约束，既保证了TRPO的算法稳定性，又降低了计算成本。使用了截断函数的PPO的目标函数可以表示为：

$$\mathcal{L}^{\mathrm{CLIP}}(\theta)=\hat{\mathbb{E}}_{\pi_{\theta_t}t}\left[\min r_t(\theta)\hat{A}_{\pi_{\theta_t}t},\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_{\pi_{\theta_t}t}\right]\quad(17)$$

式中：clip(·)为截断函数，将新旧策略的变化控制在[1−ε,1+ε]内；ε 为截断常数，用来设定策略更新的范围。

图2为PPO截断函数的示意图。

如图2所示，当优势函数的估计为负值时代表

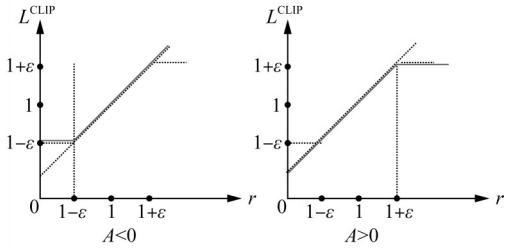

图2　PPO的clip截断函数

Fig. 2　Clip truncation function of PPO

当前的策略是消极的，则应该降低其出现的概率（以 $1-\varepsilon$ 为界限）。当优势函数的估计为正值时表示当前策略是积极的，则应在一定范围内（以 $1+\varepsilon$ 为界限）增加其概率[27]。

2.4 基于GCN-PPO算法的VVO求解流程

本文提出的GCN-PPO算法在训练时智能体基于历史运行数据和配电系统不断互动来执行MDP，在这个过程中Actor网络和Critic的网络的参数不断被更新，每回合训练结束后将网络参数保存，然后根据训练完的Actor模型来实时执行VVO。PPO-GNN算法的整体执行框架如图3所示，其离线训练的流程如表1所示。

Actor网络是将状态 s_t 映射到动作 a_t 的策略函数，其参数 θ 通常根据式（10）进行梯度更新，式中的 $\nabla_\theta J(\theta)$ 可以将策略分布调整到具有更大奖励值的方向。为了提高数据的效率，防止策略变化过大，本文引入截断函数并使用式（17）来更新参数 θ 和 θ_{old}。式（17）中的 $\hat{A}_{\pi_{\theta_t}}$ 为优势函数，其表达式为式（18）。

$$\hat{A}_{\pi_{\theta_t}} = r(s_t, a_t) + \gamma V(s_{t+1}) - V(s_t) \tag{18}$$

式（18）也代表了时序差分误差，它表示在状态 s_t 下执行动作 a_t 的优势大于所有动作的预期奖励值。由于 $r(s_t, a_t)$ 是即时奖励，式（18）可以被参数化为Critic网络来逐步更新其参数，所以优势函数的参数 μ 可以通过最小化式（19）中的 $\mathcal{L}(\mu)$ 来更新。

$$\mathcal{L}(\mu) = \mathbb{E}((V(s_t) - y_t)^2) \tag{19}$$

$$y_t = r(s_t, a_t) + \gamma V(s_{t+1}) \tag{20}$$

在训练开始时，Actor网络的参数 θ 和 θ_{old} 以及Critic网络的参数 μ 被随机初始化，其中旧策略的参数 θ_{old} 从新策略那里复制得来。在训练过程中，本文将智能体和电力系统环境互动一天作为一个回合 T，由于本文使用的历史数据的间隔是3 min，因此将每次互动作为一个时间步 t（T=480 t）。在每一个回合 T 内，智能体先和环境互动480步来形成一组旧策略，在每一更新步 t 内，Actor根据当前的状态 s_t 做出相应的动作 a_t，然后得到一个奖励 r_t 并将状态转移到 s_{t+1}。然后使用式（18）计算优势估计值，当Actor完成 T 步的交互时，Actor网络的参数 θ 通过式（21）来更新。

$$\theta \leftarrow \theta - \varphi_\theta \nabla_\theta \mathcal{L}(\theta) \tag{21}$$

表1 GCN-PPO离线训练流程

Tab. 1 Flowchart of GCN-PPO offline training

算法：GCN-PPO算法的离线训练流程

随机初始化Actor网络参数 θ 和Critic网络参数 μ
for T=1:500 **do**
  for t=1:480 **do**
    获取当前时刻的 s_t，根据初始化的策略模型 π 选择并执行 a_t，获得即时奖励 r_t 和下一时刻的状态 s_{t+1}；
    根据式（16）计算优势估计值；
    end for
  收集一个 T 内智能体和电力系统环境交互的经验，并存入一个缓存池内，然后对各网络的参数进行更新；
  for 每一个更新步 **do**
  从缓存池中取出历史经验作为更新样本；
  根据式（19）更新Actor网络的参数 θ；
  根据式（20）更新Critic网络的参数 μ；
  将Actor网络的参数赋给“旧Actor网络”：$\theta_{\text{old}} \leftarrow \theta$；
  end for
end for

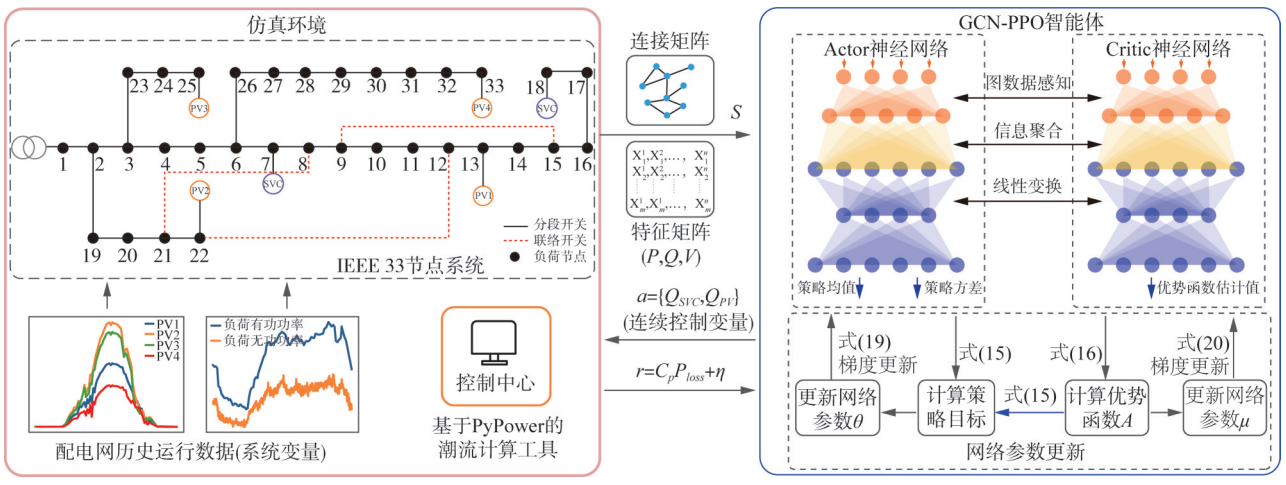

图3 基于GCN-PPO的配电网VVO流程

Fig. 3 Process of VVO in distribution network based on GCN-PPO

式中φ_θ为 Actor 网络的学习率。同时本文可以多次使用T步内收集的历史经验来更新参数θ。同样，Critic 网络的参数可以通过式(22)更新。

$$\mu \leftarrow \mu - \varphi_\mu \nabla_\mu \mathcal{L}(\mu) \tag{22}$$

式中φ_μ为 Critic 网络的学习率。每一个回合T更新完后将策略网络的参数赋给旧策略：$\theta_{\text{old}} \leftarrow \theta$。

3　算例分析

3.1　仿真环境设计

本文使用图 3 中的 IEEE 33 节点测试系统验证了所提方法的有效性[28]，该测试系统有 33 个节点，37 条线路，其中包括 32 条常闭线路和 5 条常开线路。在仿真阶段，测试系统使用 Python 中的 PyPower 工具包来进行潮流计算，所有 DRL 算法的设计和训练由 PyTorch 工具包在配备 16 GB 内存和 2.50 GHz Intel(R) Core(TM) i7-11700 的计算机上完成。

测试系统中的 4 台 PV 分别被安装在节点 13、22、25、33 处，其装机容量分别为 0.8 MW、1.2 MW、1.5 MW 和 0.5 MW。2 台 SVC 分别被安装在 7 和 18 节点，其无功功率补偿容量都是 1 Mvar。本文使用云南省某地区两年的分布式 PV 出力实测数据和负荷实测数据来训练 DRL 智能体，PV 和负荷数据的出力间隔为 3 min，因此智能体将从 240 000 (500×480)组数据中捕获问题的不确定性以及环境的动态特性[29]。同时选取夏至日和冬至日两天的数据作为测试集来观测训练的效果，这两个典型日的 PV 和负荷数据如图 4 所示。

在离线训练阶段，DRL 智能体被训练了 500 回合，每回合训练 480 步，电压限制设置为 0.97~1.07 p.u.，若电压越限则受到相应惩罚。若在训练过程中潮流不收敛，则返回终止指令。Actor 网络和 Critic 网络的超参数以及奖励函数的权重设计如表 2 所示。

3.2　DRL 方法的有效性分析

为了验证所提 DRL 方法求解 VVO 任务时的有效性，本文将 PPO 算法作为 DRL 方法的基线与所提 GCN-PPO 算法进行对比。通过比较两种 DRL 算法训练过程中的奖励值和神经网络的 loss 来判断其稳定性和有效性。两种算法的 Actor 网络和 Critic 网络使用同样的设计和超参数，并在相同的 IEEE-33

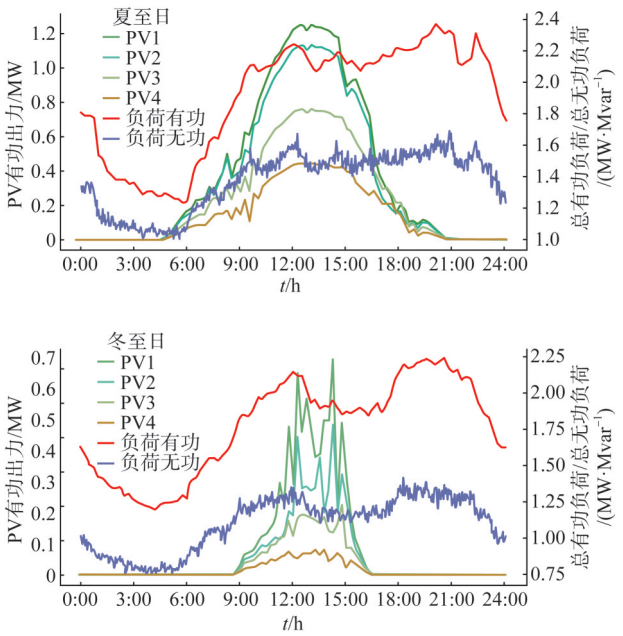

图 4　两个典型日的 PV 和负荷数据

Fig. 4　PV and load data on two typical day

表 2　参数设置

Tab. 2　Parameter settings

参数名称	数值
MLP 隐藏层尺寸	(256,256,256)
神经网络激活函数	ReLU
折扣因子γ	0.99
Actor 网络的学习率φ_θ	2×10^{-4}
Critic 网络的学习率φ_μ	2×10^{-4}
裁剪因子ε	0.2
训练回合数	500
每回合最大步数	480
网损惩罚权重	−10
电压越限惩罚权重	−200

节点测试系统上训练 500 个回合，训练的回合累计奖励值和电压越限次数如图 5 所示。由于基于 DRL 算法的随机性，本文使用不同的随机种子对每个算法进行 3 次仿真实验，训练结果的平均值和误差界在图 5 中以实线和填充区表示。训练过程中 GCN-PPO 算法的 Critic 和 Actor 网络的损失(loss)值如图 6 所示，其中 Critic 的 loss 代表网络输出的 Q 值和式(20)中标签 y 的差值，Actor 的 loss 是 Critic 网络给出来的估计值。

在训练刚开始时，随机初始化的参数使两种 DRL 智能体均表现出较低的性能，具体表现为电压越限严重，此时智能体所做的动作是无效的，从图

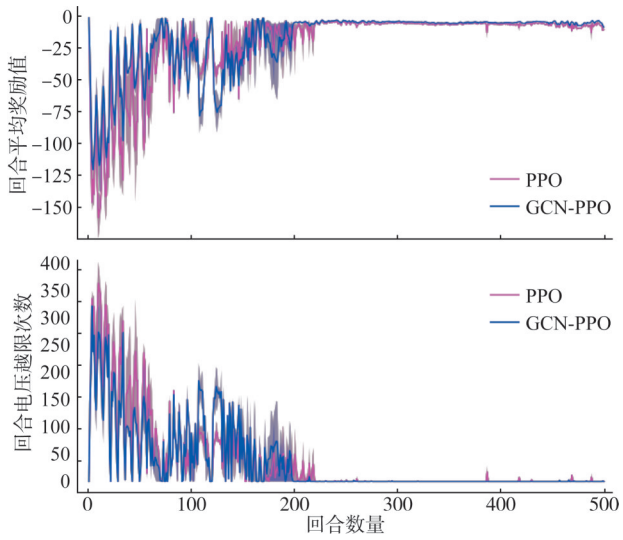

图 5 训练过程中的奖励值和电压越限次数

Fig. 5 Reward value and number of voltage limit violation during training

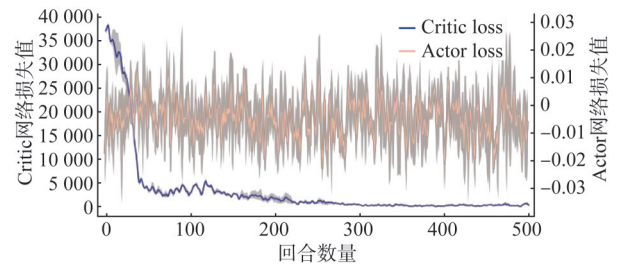

图 6 训练过程中 Critic 和 Actor 网络的 loss 值

Fig. 6 Loss values of critic and actor networks during training

6 中也可以看出 Critic 网络输出的 Q 值和标签 y 相差较大。随着训练的进行，智能体在奖励值的指导下开始从经验中学习正确的策略，同时根据损失函数提供的梯度，Critic 网络的参数不断被优化，loss 也不断减小。在训练达到 200 回合后，智能体已经可以做出正确动作，此时奖励值和 loss 开始收敛，电压每回合越限的次数也趋近于 0。训练结束后两种算法均收敛到较高的奖励值，并且不再出现电压越限的情况。从图 5 中可以看出，GCN-PPO 收敛后的性能要略优于 PPO 算法，由于所提算法可以更好的感知配电网的拓扑结构和潮流的空间分布情况，智能体可以做出更精准的调压动作，所以在训练达到稳定后奖励值更高。但由于 GCN-PPO 算法具有更复杂的神经网络结构，因此收敛过程要慢与 PPO 算法。

3.3 各种 VVO 方法性能对比

DRL 方法在执行 VVO 任务时的决策过程就是将训练好的智能体（Actor 神经网络）保存下来并部署到现场，这个过程也可以被称为测试阶段。为了证明所提 GCN-PPO 算法在执行 VVO 任务时的优势，本文将训练好的 GCN-PPO 算法和 PPO 算法在测试集上进行对比，同时在已知系统不确定性量的情况下（测试集中 PV 发电和负荷用电情况），与基于模型的基于启发式的 PSO 算法[30]和 OPF 算法和[31]进行数值比较。

图 7 展示了上述 4 种控制算法分别在夏至日和冬至日测试集上的有功网损对比。

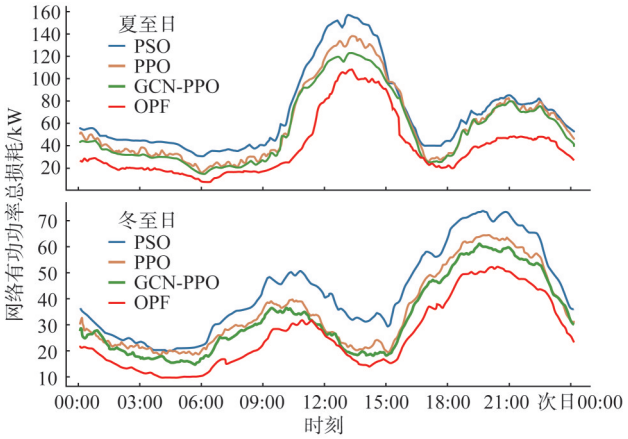

图 7 4 种方法在测试集上的网损

Fig. 7 Network losses of 4 methods on the test sets

由于 OPF 算法是在假定知道系统参数和不确定性变量的情况下进行的最优化计算[2]，因此求得的解是最优的，但是 OPF 是一种确定性优化算法，并不能处理优化问题中的不确定性，本文只是将 OPF 求得的最优解作为可供其他算法对比的最优方案。从图 7 中可以看出，本文所提的 GCN-PPO 算法表现出与 OPF 相似的性能。GCN-PPO 算法在夏至日测试集上降低网损的性能相比于 PSO 和 PPO 算法分别提升了 5.2% 和 17.6%，在冬至日测试集上的性能分别提升了 4.9% 和 21.3%。

图 8 为节点 33 分别在夏至日和冬至日测试集上一天内的电压波动情况，图 8 中展示了 4 种控制方法和不加任何控制措施（None）的情况下的电压幅值。从图 8 中可以看出在夏至日 33 节点在没有任何控制措施时中午和晚上均存在电压越限（红色虚线）的情况，这是由于中午 PV 出力过大导致的过电压和晚上用电负荷过大导致的欠电压，而冬至日也会在晚上出现欠电压的情况。4 种控制算法均可以有

效缓解电压越限，其中PSO可以很好地解决欠电压问题，但是无法解决过电压问题，OPF在优化求解时定义了约束条件，因此可以确保满足电压约束。从图8中可以看出本文所提GCN-PPO算法的电压优化效果要明显优于PSO和PPO算法。

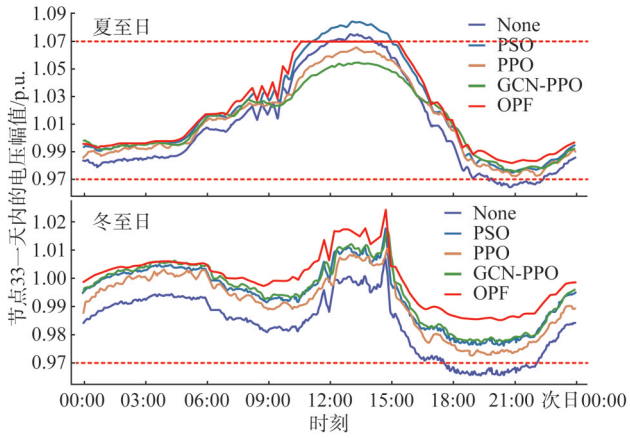

图8　4种控制方法对节点33的电压幅值优化对比

Fig. 8　Optimization comparison of 4 control methods for voltage magnitudes at node 33

图9展示了4种优化方法在冬至日测试集晚上20：00时对各节点电压的优化情况，可以看出不施加任何控制措施时许多配电网末端的节点会出现欠电压的情况。4种优化方法均可以解决节点欠电压的情况其中本文所提GCN-PPO算法相比于PSO算法和PPO算法效果分别提升了22.35%和11.15%，这是由于GCN-PPO可以更好地感知配电网的潮流分布，进而给出更佳的调压策略。

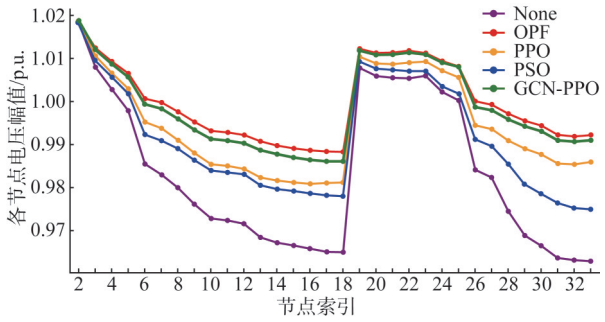

图9　4种方法在20:00时刻对各节点电压的优化对比

Fig. 9　Optimization comparison of 4 methods for each node voltages at 20:00

表3展示了4种优化算法每一个决策步求解策略的时间。PSO算法在每个决策开始时需要多次进行潮流计算并更新粒子参数寻找最优解，而OPF算法也需要依靠商业求解器在每个决策步对VVO模型不断的迭代求解，它们的求解速度取决于问题的规模和求解器的计算效率，在本文使用的33节点测试系统中求解过程在秒级可以完成。相比而言，本文所提GCN-PPO算法将训练好的智能体看作一个附带神经网络模型的控制器，因此在决策时只需对其输入当前时刻的状态便可以得到一组策略动作，求解时间可以达到毫秒级，相比于基于模型的方法求解速度快了2~3个数量级。从表中可以看出，所提GCN-PPO算法由于具有更复杂的网络结构，求解时间比PPO算法慢1.56ms，但是为了更高的求解质量这些时间是可以接受的。

表3　单步决策时间对比

Tab. 3　Time comparison of single-step decision

算法	平均耗时
PSO	3.7 s
OPF	1.67 s
PPO	6.79 ms
GCN-PPO	8.35 ms

3.4　所提方法可扩展性和鲁棒性分析

为了验证所提算法的控制策略在配电网拓扑发生变化时的可扩展性，本文分别在春夏秋冬4个季节选取了1d的PV和负荷运行数据，并设置了3种配电网拓扑变化的场景，来对比GCN-PPO、PPO、PSO和OPF 4种方法的控制策略在不同场景上的可扩展性。

为了模拟故障维修时的情景，本文将闭合联络开关(图3中红色的虚线)和打开分段开关(图3中黑色的实线)作为一组拓扑变换的方式。并根据拓扑变换的程度设计了3个场景，1)场景1是闭合12和22节点间的联络开关，打开10和11节点的分段开关；2)场景2是闭合12和22、8和21、25和29节点之间的联络开关，然后打开10和11、4和5、6和26节点之间的分段开关；3)闭合12和22、8和21、25和29、9和15、18和33节点之间的联络开关，然后打开10和11、4和5、6和26、8和9、30和31节点之间的分段开关。

图10展示了3个场景下4种方法在所选测试集上的累计电压越限节点数。从图10中可以看出，当配电网拓扑结构发生变化时，OPF和PSO算法在原始拓扑上所求得的控制策略并不具备扩展到新拓

扑的能力。而普通的PPO智能体在训练过程中由于没有考虑拓扑的连接关系，随着配电网拓扑的变化程度不断加大，其电压越限率也在不断增大。本文所提GCN-PPO智能体在拓扑变化较大的情况下依然具备良好的电压控制性能，所求策略的可扩展性要优于另外3种方法。

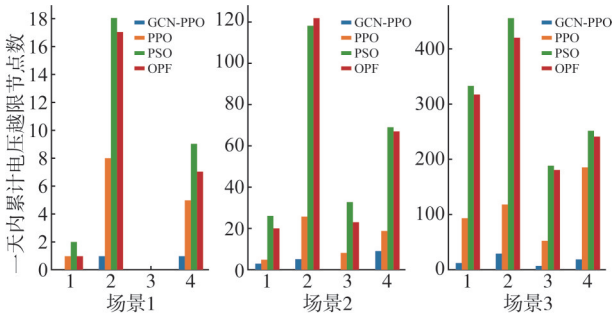

图10 3个场景下的累计电压越限节点数对比

Fig. 10 Comparison of the cumulative number of voltage limit violation nodes in three scenarios

在电力系统中，由于传感器通讯故障导致的测量数据丢失的情况并不少见，因此所收集的历史运行数据可能存在数据丢失的情况，这对于数据驱动的DRL方法来说是非常不利的，本文对比了在数据丢失时GCN-PPO和PPO两种DRL方法的鲁棒性。

为了模拟测量数据丢失的情况，本文随机选择配电网的一些节点，在测试期间将其电压设置为零。这导致PPO算法输入的状态向量的某些值零，而GCN-PPO算法输入的图数据中的某些节点特征为零。同时设置了3组实验场景，分别代表25%、50%、75%数量的节点电压测量值丢失。表4展示了所提GCN-PPO算法和PPO算法在3种场景下相比没有丢失数据时的奖励值的减小量，该减小量为在不同的随机种子下3次仿真实验所得结果的平均值。

从表4中可以观察到尽管2种算法的性能都受到了数据丢失的影响，但是与GCN-PPO相比，数据丢失对PPO算法的影响要严重得多，且随着数据丢失量的加大这种影响也会加大，在最坏的情况下(数据丢失75%时)PPO的性能比GCN-PPO差将近3.6倍。因此，在测量数据丢失的情况时GCN-PPO比PPO具有更好的鲁棒性。

在实际情况中，电网调度人员通常使用相邻节点电压的平均值补全缺失的数据，这种做法和GCN-PPO中使用的图卷积层中消息传递的原理相似。但本文认为GCN-PPO算法是实现这种鲁棒性的更具原则的方式，因为GCN-PPO学习了邻接值的(可能是非线性)聚合策略，而不是需要人工手动制定的策略。

表4 数据丢失对算法性能的影响

Tab. 4 Impacts of data losses on algorithm performance

算法	GCN-PPO	PPO
25%数据丢失	0.93	1.87
50%数据丢失	1.76	5.65
75%数据丢失	4.42	15.89

4 结论

本文提出了一种基于GCN-PPO算法的配电网电压无功优化方法，通过在DRL模型中嵌入GCN使智能体在训练时可以感知配电网的图数据，因此在配电网拓扑变化时智能体具有良好的扩展性能，同时在测量数据丢失时GCN-PPO算法也能表现出良好的鲁棒性。通过在改进的IEEE33节点系统上的数值实验发现，本文所提方法在训练阶段可以收敛到很好的结果，能够在确保电压不越限的情况下有效地降低网络的有功损耗。

在测试集上的验证表明，所提方法具有和基于模型的OPF方法相似的性能，且在夏至日和冬至日测试集上降低网损的性能比PSO及PPO算法分别提升了5.2%、17.6%和4.9%、21.3%。对于节点电压的优化，所提GCN-PPO算法要明优于PSO和PPO方法，其中对于冬季欠电压的优化效果相比于PSO和PPO提升了22.35%和11.15%。在求解速度方面，比OPF和PSO方法快了2~3个数量级，在每个决策步都可以实现毫秒级求解。

最后，通过实验发现本文所提GCN-PPO算法在拓扑变化较大的情况下仍具备良好的电压控制性能，所求策略的可扩展性要优于上述3种方法。同时在数据丢失较多的情况下GCN-PPO算法的性能也会高于PPO算法3.6倍，具备良好的鲁棒性。

参考文献

[1] 尹昌洁，权楠，苏凯，等. 我国分布式能源发展现状及展望[J]. 分布式能源，2022，7(2)：1－7.
YIN Changjie，QUAN Nan，SU Kai，et al. Status and outlook of distributed energy development in China [J]. Distributed

Energy，2022，7(2)：1－7.

[2] 赵乐，许凌，张梦瑶，等.考虑非完美通信的高比例光伏配电网模糊电压控制策略[J].广东电力，2023，36(12)：9－16.
ZHAO Le，XU Ling，ZHANG Mengyao，et al. Fuzzy voltage control strategy for high proportion photovoltaic distribution network considering nonperfect communication[J]. Guangdong Electric Power，2023，36(12)：9－16.

[3] VALVERDE G，CUTSEM T V. Model predictive control of voltages in active distribution networks[J]. IEEE Transactions on Smart Grid，2013，4(4)：2152－2161.

[4] XU Y，DONG Z Y，ZHANG R，et al. Multi-timescale coordinated voltage/var control of high renewable-penetrated distribution systems[J]. IEEE Transactions on Power Systems，2017，32(6)：4398－4408.

[5] 彭宇文，李瑞，周永旺，等.基于IGDT和PSO-SA的综合能源系统鲁棒优化调度方法[J].广东电力，2023，36(9)：60－71.
PENG Yuwen，LI Rui，ZHOU Yongwang，et al. Robust optimal scheduling method for integrated energy system based on IGDT and PSO-S[J]. Guangdong Electric Power，2023，36(9)：60－71.

[6] LI P S，ZHANG C，WU Z J，et al. Distributed adaptive robust voltage/var control with network partition in active distribution networks[J]. IEEE Transactions on Smart Grid，2020，11(3)：2245－2256.

[7] 李程昊，蒋芒，姚德贵，等.高比例可再生能源电力系统的日前-日内两级无功功率滚动调度方法[J].南方电网技术，2022，16(6)：94－103.
LI Chenghao，JIANG Mang，YAO Degui，et al. Day-ahead and intra-day two-stage reactive power rolling dispatch method for power systems with high renewable power penetration[J]. Southern Power System Technology，2022，16(6)：94－103.

[8] ZHAO B，XU Z C，XU C，et al. Network partition-based zonal voltage control for distribution networks with distributed PV systems[J]. IEEE Transactions on Smart Grid，2018，9(5)：4087－4098.

[9] 吴国沛，王武，张勇军，等.含光储系统的增量配电网时段解耦动态拓展无功优化[J].电力系统保护与控制，2019，47(9)：173－179.
WU Guopei，WANG Wu，ZHANG Yongjun，et al. Time decoupled dynamic extended reactive power optimization in incremental distribution network with photovoltaic-energy storage hybrid system[J]，Power System Protection and Control，2019，47(9)：173－179.

[10] 李子健，郭佩乾，马宁宁，等.融合双重策略粒子群算法的分布式电源配网无功优化[J].南方电网技术，2022，16(6)：14－22，81.
LI Zijian，GUO Peiqian，MA Ningning，et al. Reactive power optimization for distribution system with DG by particle swarm optimization algorithm integrating dual strategies[J]. Southern Power System Technology，2022，16(6)：14－22，81.

[11] 黄伟，刘斯亮，王武，等.长时间尺度下计及光伏不确定性的配电网无功优化调度[J]电力系统自动化，2018，42(5)：154－162.
HUANG Wei，LIU Siliang，WANG Wu，et al. Optimal reactive power dispatch with long-time scale in distribution network considering uncertainty of photovoltaic [J]. Automation of Electric Power System，2018，42(5)：154－162.

[12] WU M H，HONG L C，WANG Y F，et al. Volt-var control for distribution networks with high penetration of DGs：an overview[J]. The Electricity Journal，2022，35(5)：107130.

[13] ZHANG B，GAO Y，WANG L，et al. BO-XGBoost-based voltage/var optimization for distribution network considering the LCOE of PV system[J]. IET Renewable Power Generation，2024，18(3)：502－514.

[14] 李琦，乔颖，张宇精.配电网持续无功优化的深度强化学习方法[J].电网技术，2020，44(4)：1473－1480.
LI Qi，QIAO Ying，ZHANG Yujing，et al. Continuous reactive power optimization of distribution network using deep reinforcement learning[J]. Power System Technology，2020，44(4)：1473－1480.

[15] CAO D，ZHAO J B，HU J X，et al. Physics-Informed Graphical Representation-Enabled Deep Reinforcement Learning for Robust Distribution System Voltage Control[J]. IEEE Transactions Smart Grid，2024，15(1)：233－246.

[16] CAO D，HU W H，XU X，et al. Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices [J]. Journal of Modern Power Systems and Clean Energy，2021，9(5)：1101－1110.

[17] 胡丹尔，彭勇刚，韦巍，等.多时间尺度的配电网深度强化学习无功优化策略[J].中国电机工程学报，2022，42(14)：5034－5045.
HU Daner，PENG Yonggang，WEI Wei，et al. Multi-timescale deep reinforcement learning for reactive power optimization of distribution network[J]. Proceedings of the CSEE，2022，42(14)：5034－5045.

[18] WANG S Y，DUAN J J，SHI D，et al. A data-driven multi-agent autonomous voltage control framework using deep reinforcement learning [J]. IEEE Transactions on Power Systems，2020，35(6)：4644－4654.

[19] CAO D，ZHAO J B，HU W H，et al. Data-driven multi-agent deep reinforcement learning for distribution system decentralized voltage control with high penetration of PVs[J]. IEEE Transactions on Smart Grid，2021，12(5)：4137－4150.

[20] 王铮澄，周艳真，郭庆来，等.考虑电力系统拓扑变化的消息传递图神经网络暂态稳定评估[J].中国电机工程学报，2021，41(7)：2341－2350.
WANG Zhengcheng，ZHOU Yanzhen，GUO Qinglai，et al. Transient stability assessment of power system considering topological change：a message passing neural network-based approach[J]. Proceedings of the CSEE，2021，41(7)：2341－2350.

[21] ZHAO T Q，WANG J H. Learning sequential distribution system restoration via graph-reinforcement learning[J]. IEEE Transactions on Power Systems，2022，37(2)：1601－1611.

[22] LEE X Y，SARKAR S，WANG Y. A graph policy network approach for volt-var control in power distribution systems[J]. Applied Energy，2022，(323)：119530.

[23] HONG L C，WU M H，WANG Y F，et al. MADRL-based DSO-customer coordinated bi-Level volt/var optimization method for power distribution networks [J]. IEEE Transactions on Sustainable Energy，2024，15(3)：1834－1846.

[24] SUTTON R S，BARTO A G. Reinforcement learning：an introduction[M].2nd ed.Massachusetts：MIT press，2018.

[25] UTTON R S，BARTO A G. Reinforcement Learning：An Introduction [J]. IEEE Transactions on Neural Networks，1998，9(5)：1054.

[26] 刘浩宇，刘挺坚，刘友波，等. 基于图卷积神经网络的直流送端系统暂态过电压评估[J]. 电力系统保护与控制，2023，51(23)：71－81.
LIU Haoyu，LIU Tingjian，LIU Youbo，et al. A method for evaluating transient overvoltage of an HVDC sending-end system based on a graph convolutional network [J]. Power System Protection and Control，2023，51(23)：71－81.
[27] 祁向龙，陈健，赵浩然，等. 多时间尺度协同的配电网分层深度强化学习电压控制策略[J]. 电力系统保护与控制，2024，52(18)：53－64.
QI Xianglong，CHEN Jian，ZHAO Haoran，et al. Multi-time scale cooperative voltage control strategy of a distribution network based on hierarchical deep reinforcement learning [J]. Power System Protection and Control，2024，52(18)：53－64.
[28] 黄东晋，蒋晨凤，韩凯丽. 基于深度强化学习的三维路径规划算法[J]. 计算机工程与应用，2020，56(15)：30－36.
HUANG Dongjin，JIANG Chenfeng，HAN Kaili. 3D path planning algorithm based on deep reinforcement learning [J]. Computer Engineering and Applications，2020，56(15)：30－36.
[29] 张波，高远，李铁成，等. 考虑光伏电源可靠性的新能源配电网数据驱动无功电压优化控制[J]. 中国电机工程学报，2024，44(15)：5934－5947.
ZHANG Bo，GAO Yuan，LI Tiecheng，et al. Data-driven voltage/var optimization control of active distribution network considering the reliability of photovoltaic power supply [J]. Proceedings of the CSEE，2024，44(15)：5934－5947.
[30] WANG Y，ZHAO T Y，JU C Q，et al. Two-level distributed volt/var control using aggregated PV inverters in distribution networks [J]. IEEE Transactions on Power Delivery，2020，35(4)：1844－1855.
[31] 韩子颜，王守相，赵倩宇，等. 计及分时电价的5G基站光储系统容量优化配置方法[J]. 中国电力，2022，55(9)：8－15.
HAN Ziyan，WANG Shouxiang，ZHAO Qianyu，et al. A capacity optimization configuration method for photovoltaic and energy storage system of 5G base station considering time-of-use electricity price [J]. Electric Power，2022，55(9)：8－15.

收稿日期：2023-04-10；**网络首发日期**：2024-02-20
作者简介：
朱涛(1978)，男，高级工程师(教授级)，博士，研究方向为电力系统调度运行管理、调度自动化，taozh78@163.com；
海迪(1992)，女，工程师，硕士，研究方向为配网调度运行控制、配电自动化，1939543979@qq.com；
吴明贺(1997)，男，通信作者，博士研究生，研究方向为深度强化学习在电力系统中的应用、分布式能源能量管理，mh_wu@seu.edu.cn。

(上接第57页 Continued from Page 57)

[32] XAVIER Á S，QIU F，AHMED S. Learning to solve large-scale security-constrained unit commitment problems [J]. INFORMS Journal on Computing，2020，33(2)：739－756.
[33] PINEDA S，MORALES J M，JIMÉNEZ-CORDERO A. Data-driven screening of network constraints for unit commitment [J]. IEEE Transactions on Power Systems，2020，35(5)：3695－3705.
[34] 沈海平，陈铭，钱磊，等. 计及电转气耦合的电-气互联系统机组组合线性模型研究[J]. 电力系统保护与控制，2019，47(8)：34－41.
SHEN Haiping，CHEN Ming，QIAN Lei，et al. Linear model research of unit commitment for integrated electricity and natural-gas systems considering power-to-gas coupling [J]. Power System Protection and Control，2019，47(8)：34－41.
[35] 赵文猛，刘明波. 多区域电力系统分散式动态经济调度的割平面一致性算法[J]. 中国电机工程学报，2016，36(21)：5806－5815，6023.
ZHAO Wenmeng，LIU Mingbo. Cutting plane consensus algorithm for decentralized dynamic economic dispatch in multi-area power systems [J]. Proceedings of the CSEE，2016，36(21)：5806－5815，6023.

收稿日期：2023-05-25；**网络首发日期**：2024-03-19
作者简介：
陈子瑞(1998)，男，通信作者，硕士研究生，研究方向为电力系统智能化调度和运行控制，751357830@qq.com；
刘明波(1964)，男，教授，博士生导师，博士，研究方向为电力系统调度与电力市场，epmbliu@scut.edu.cn；
曾贵华(1998)，男，硕士研究生，研究方向为电力系统智能化调度和综合能源系统优化运行，185071364@qq.com。