文章编号：1000-3673（2022）07-2547-08　　中图分类号：TM 721　　文献标志码：A　　学科代码：470 40

基于深度强化学习组合优化的配电网拓扑控制研究

闫冬[1]，彭国政[1]，高海龙[2]，陈盛[1]，周钰山[2]
（1．中国电力科学研究院有限公司，北京市 海淀区 100192；
2．国网江苏省电力有限公司徐州供电分公司，江苏省 徐州市 221000）

Research on Distribution Network Topology Control Based on Deep Reinforcement Learning Combinatorial Optimization

YAN Dong[1], PENG Guozheng[1], GAO Hailong[2], CHEN Sheng[1], ZHOU Yushan[2]
(1. China Electric Power Research Institute, Haidian District, Beijing 100192, China;
2. State Grid Jiangsu Xuzhou Power Supply Company, Xuzhou 221000, Jiangsu Province, China)

ABSTRACT: This paper studies the topology control strategy of distribution network fault recovery based on deep reinforcement learning. First, design the distribution network topology state representation and decision-making action rules to support the combined optimization solution. Secondly, use the improved pointer network structure and deep reinforcement learning algorithm to achieve model self-learning and end-to-end calculation suitable for multiple types of failure recovery strategies. Finally, by improving the mask mechanism to reduce the complexity of exploration and solving, and effectively improve the efficiency of training and learning. By randomly setting fault combinations on the preset lines, the effectiveness of the improved mechanism and model proposed in this paper is verified on the single and mixed initial state sample sets, which provides an effective reference for the application of deep learning technology in the optimization of distribution network operation mode.

KEY WORDS: deep reinforcement learning; topology control; combination optimization; pointer networks

摘要：基于深度强化学习的配电网故障恢复拓扑控制策略，文章首先设计配电网拓扑状态表征和决策动作规则支撑组合优化求解；其次，利用改进指针网络结构配合深度强化学习算法实现适用于多类故障恢复策略的模型自学习和端到端计算；最后，通过改进掩码机制降低探索求解复杂度进而提升训练学习效率。通过在预设条线路上随机设置故障组合，在单一和混合初始状态样本集上验证文章提出的改进机制和模型计算有效性，以期为深度学习技术在配电网运行方式优化研究提供有效参考。

关键词：深度强化学习；拓扑控制；组合优化；指针网络
DOI：10.13335/j.1000-3673.pst.2021.1388

基金项目：国家电网有限公司总部科技项目(5700-202018266A-0-0-00)。
Project Supported by the Project of State Grid Company (5700-202018266A-0-0-00).

0　引言

配网侧安全运行是保证电网稳定的核心[1]，而依靠配网拓扑的灵活性改善或修正配网运行方式成为重要的控制手段，尤其是故障后的供电快速恢复是保证其供电可靠性的关键。传统的配电网拓扑控制相关研究包括运行方式重构[2]，考虑故障后的负荷转供策略[3-4]等。拓扑优化问题在研究中往往需要将各类控制开关状态定义为 0–1 二元变量，而整数的引入及系统模型复杂性造成传统优化的应用受限，故常基于如改进粒子群[5]、量子人工蜂群[6]等的启发式优化算法进行求解。以上研究有效地实现了网损最小化、运行可靠性最大化等配网优化运行目标，在分布式新能源接入情况下也可有效计算[7]。

人工智能技术的出现为配电网拓扑控制提供新思路，有助于实现从运行特征到网络控制策略的端到端决策，从而突破传统优化计算的实时性瓶颈，这其中深度强化学习技术作为机器学习优化决策的核心技术受到广泛关注，其核心是深度神经网络的强大特征提取能力，以及强化学习技术提供的策略提升能力。文献[8]在随机新能源及动态负荷接入的主网场景中利用包含拓扑控制策略的决策集，实现了网络稳定运行，其核心思路是基于常规运行经验对深度神经网络进行预训练，配合对动作策略集的筛选降维，在测试中使用深度强化学习对网络进行策略提升；文献[9]利用图神经网络提取电网拓

扑变化及运行物理信息特征，有效地建立控制策略效果与状态表征之间的联系，从而更好地支撑深度强化学习的交互更新。为了达到以固定参数模型实现动态多场景决策的目标，目前研究通常将拓扑控制与其他配套措施设备调整组合实施，以模仿学习而非纯粹智能体自学习方式引导模型更新。这样的目的一是为了解决问题规模和信息维度大的问题，更重要的是突破以单步决策解决优化问题造成的探索难题。深度强化学习应用需要严格稳定的仿真交互环境，且求解问题的马尔科夫决策建模完整性同样影响策略训练的效果[10]。但这些核心因素的构架在迁移至配电网拓扑优化问题上遇到困难，一方面拓扑的改变使环境变得不稳定，导致动作意义发生变化；另一方面网络运行状态优劣评价依靠组合效果而非交互的长期收益，很难从时序决策的角度评价拓扑在形成及改变过程中的单个时间断面内的具体价值，以单步决策建模进而使用优化算法求解显得更合理。

为了突破从优化到时序决策转化的困境，近年来深度强化学习在组合优化方向[11]的研究提供了解决思路，从实现旅行商路径优化问题[12]到车辆路径优化问题[13]再到更复杂的电动车路径优化问题[14]等，基于指针网络(pointer-networks)[15]的深度强化学习方法成功地实现了组合优化端到端计算的突破。文献[16]已针对该技术在电力系统中的应用展开研究，利用节点—支路关联矩阵构建组合元素空间，基于改进指针网络实现适用于多场景输电网网架的线路投资成本优化计算，相较传统方法在运算效率和模型灵活性上均取得较大提升。将该技术应用于同样以线路状态为核心状态表征及控制变量的配网拓扑控制中，一方面决策空间从线路选址的单向增加扩展为闭合和断开 2 类动作，复杂度更高；其次，拓扑初始运行状态多样，引入故障线路后的状态信息更加复杂，难以依靠静态关联矩阵进行描述；最后，输电网网架规划计算更注重整体结构的长期收益，在单步线路选址决策间没有动态约束限制，应用于故障状态下的配电网恢复运行中，尤其是面向线路状态改变对后续决策形成约束的问题时，将导致决策频繁越限无法有效，需要进一步改进组合元素选取与网络参数更新的交互策略，以保证优化决策有效性。本文针对以上问题，基于深度强化学习组合优化研究的先进方法和理念，探索其在配电网拓扑故障状态恢复的控制应用方法，实现适用于多类故障分布状态下的恢复运行策略神经网络自学习和端到端控制策略计算，为提升配电网自动化的人工智能技术应用进行探索尝试。

1 问题描述和建模

本文主要探讨深度强化学习技术应用于配电网拓扑控制，实现针对故障线路拓扑自适应调整和供电恢复的策略生成，以提升供电可靠性和电网运行自主性，本质问题是通过限制故障或计划线路运行状态，在可行开关状态中选择目标函数最优的组合。由于可控开关有限，且 0–1 状态不可同时取得，故其求解问题可建模为带约束的组合优化问题。本节主要介绍参量定义、问题建模与模型建模。

1.1 参量定义

用于描述问题建模和模型建模的变量含义如表 1 所示。

表 1 参量符号及定义

Table 1 Parameter symbols and definitions

符号	含义
G	配电网拓扑结构
D	配电网 G 内的负荷节点
$\|D\|$	配电网 G 内的负荷节点数目
V	配电网 G 节点间的电气连接边
$\|V\|$	配电网 G 节点间的电气连接边数
Φ	决策元素完全集
Ω	已选择的元素构成的集合
N	集合 V 内元素的编号，取值为 0~n，其中 n=$\|V\|$–1
i	集合 D 中的元素
j	集合 V 中的元素
U_i	负荷节点 i 的电压值
I_j	流过线路 j 的电流值
$\gamma_i(G)$	拓扑 G 内负荷节点的带电状态，带电时取值为 1，反之为 0
α_j/α'_j	决策改变前/后线路的开关状态，闭合时取值为 1，反之为 0

1.2 问题建模

首先介绍作为神经网络输入的组合决策完全集构建方法。拓扑内每条线路对应的开关具有闭合、断开和故障 3 种状态，其中故障状态定义为不可恢复的断开状态。对于每个开关，使用编号及对应的状态变量可构建可行解完全集。为保证状态表达形式统一，将每个正常可控开关表达为 1 组互斥元素的排列，即：

$$\{(N,\alpha_N),(N,\bar{\alpha}_N)\} \tag{1}$$

式中：元素的顺序为开关当前状态 α_N 在前，互斥状态 $\bar{\alpha}_N$ 在后。对于故障开关，其状态不存在互斥元素，以 2 项相同元素形成占位，即：

$$\{(N,0),(N,0)\} \tag{2}$$

基于以上建模方式，决策元素完全集内描述拓扑状态的元素共 2|V|个。对于每个被选到的元素 x，其意义变化为对原拓扑 G 的 1 种操作，即将对应编号 N 的开关调整至元素对应的状态 α 上。通过此类

决策序列得到的新拓扑 G'为

$$G' = x_n * x_{n-1} * \cdots * x_0 * G \tag{3}$$

式中*表示拓扑操作运算。问题的求解过程可被抽象成无放回的抽取问题，即在每个时刻选择 1 个开关及其状态，对应将拓扑操作变化，循环直到全部可控开关遍历完毕。由于开关状态唯一，故对于同 1 个编号 N 仅能进行 1 次操作，故每次抽取会从完全集中去除 2 个元素，通常需要进行共$|V|$次抽取来形成决策组合。考虑到面对不同故障分布情况时，不同开关的作用各不相同，在决策优先级上存在差异；且实际操作时往往要求操作数最少，故选取顺序直接影响决策步数及优化质量。

由于各开关本身具有拓扑关联性，进行求解需作如下假设：1）网络拓扑固定，保持不变的量包括各节点的可能连接方式，各电源的连接节点编号，开关位置及编号，以上各要素保证了每个开关状态作用的统一性，避免由于拓扑关系的改变造成变量描述的物理意义改变；2）不考虑功率缺额情况，即拓扑内连接的电源可设为无限大系统，足够支撑全部负荷要求；3）故障线路的设置不能造成运行孤岛出现，即避免通过完全集内的操作无法实现优化的状态；4）拓扑内每条边均可闭合断开，保证状态描述完备性。

对于无模型强化学习智能体，各开关的拓扑关系属于隐性知识，需要通过探索学习来构建。这些知识需要在状态描述时以统一的形式表达。基于以上问题基本描述及假设构建，可以将各开关的拓扑关系用序列表示，按照固定的开关位置顺序建立完全集。

1.3　模型建模

假设网络拓扑决策前后的线路状态为

$$\begin{cases} V = \{(0,\alpha_0),(0,\bar{\alpha}_0),\ldots,(n,\alpha_n),(n,\bar{\alpha}_n)\} \\ V' = \{(0,\alpha'_0),(0,\bar{\alpha}'_0),\ldots,(n,\alpha'_n),(n,\bar{\alpha}'_n)\} \end{cases} \tag{4}$$

则表达策略供电可靠性及快速性的目标函数定义分别为

$$\max f_1 = \sum_{i\in D} \gamma_i (G' = x_t * \cdots * x_1 * G \,|\, x_t \in \Omega) \tag{5}$$

$$\min f_2 = \sum_j |\alpha'_j - \alpha_j| \tag{6}$$

考虑的约束条件包括配网线路电流限值约束和节点电压限值约束，分别为

$$I_{j\min} \le I_j \le I_{j\max} \tag{7}$$

$$U_{i\min} \le U_i \le U_{i\max} \tag{8}$$

2　算法设计与求解

本节基于问题模型进行求解算法设计，求解过程可描述为基于当前时刻决策序列循环求取模型输出概率最大的元素，即：

$$\begin{cases} y_{t+1} = \underset{y\in\Phi}{\operatorname{argmax}}\, P(y|\boldsymbol{X}_t, \boldsymbol{Y}_t) \\ \boldsymbol{X}_{t+1} = f(\boldsymbol{X}_t, y_{t+1}) \end{cases} \tag{9}$$

式中：$\boldsymbol{X}_t$ 表示在 t 时刻的决策状态空间信息输入；y 表示在 t 时刻模型的决策输出信息；$\boldsymbol{Y}_t = \{y_0,\ldots,y_t\}$ 表示到 t 时刻为止模型完成的决策输出序列；f 表示状态空间更新的状态转移函数。由于决策需要进行 T 次选择来完成，故 t=1,2,…,T。

2.1　改进的指针网络

由式(9)可知，预测变量 y 始终在集合 Φ 内，且集合 Φ 中的元素直接构成状态空间 X。研究该类问题需要将模型输出与模型输入直接建立联系，通过输出策略反馈的奖励信号及运算过程变量构建高维特征，并进一步得到量化概率模型，即完成输入序列的自映射过程。指针网络是被设计用于解决组合选取问题的一类模型，其核心是通过注意力机制[17]完成特征对应及概率计算。

2.1.1　指针网络基本结构

改进的指针网络基础逻辑结构如图 1 所示。图中状态序列信息首先输入到编码器(encoder)实现序列编码，目的是将显性信息转化为高维特征向量。编码器实现的向量嵌入一般由卷积或循环神经网络结构完成，考虑到编码器输入的决策状态特征及数据结构，本研究中采用一维卷积网络结构作为编码器提取状态空间序列包含的隐性拓扑特征。然后，将编码器输出的高维特征作为解码器(decoder)的部分输入，结合注意力机制对原状态集合进行预测，通过求解式(9)循环求得网络拓扑组合。解码器的计算方式与循环神经网络(recurrent neural network，RNN)网络预测的方式相同，本文采用门控循环单元(gate recurrent unit，GRU)[18]作为模块核心结构完成解码功能。

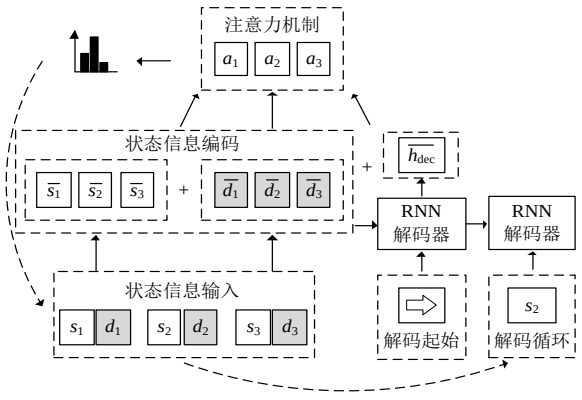

图 1　改进的指针网络基础逻辑结构

Fig. 1　Improved pointer network structure

2.1.2 注意力机制

注意力机制用于构建当前时刻决策隐藏特征对应各决策完全集元素的概率分布，隐藏特征包括解码器输出隐藏层变量特征及决策完全集编码特征。通过维护 2 个参数矩阵及激活函数构建融合 2 类特征的内容矩阵，将内容矩阵利用 softmax 函数计算归一化的概率表征，即：

$$\begin{cases} \boldsymbol{h}_t = \boldsymbol{W}_1 \tanh(\boldsymbol{W}_2[\boldsymbol{h}_{\text{enc}};\boldsymbol{h}_{\text{dec}}]) \\ \boldsymbol{a}_t = \text{softmax}(\boldsymbol{h}_t) \end{cases} \tag{10}$$

式中：$\boldsymbol{h}_{\text{enc}}$ 表示经编码器处理的决策完全集特征；$\boldsymbol{h}_{\text{dec}}$ 表示经解码器处理的上一时刻选择的状态元素信息；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 为可训练的参数矩阵；tanh 为激活函数；$\boldsymbol{a}_t$ 为对应元素状态序列的概率值向量。上一时刻的输出与当前时刻的可选元素之间的映射概率构建由 2 层注意力机制完成，分别是计算输出结果返回解码器计算隐层变量与可选元素复合状态信息共同编码的注意力内容向量，其意义为构建已完成策略的高维特征。之后，计算该内容向量与可选元素静态信息共同编码的注意力概率模型，即得到目标的映射关系。

2.1.3 动态信息嵌入

初始版本的指针网络只考虑静态状态空间的输入情况，即 $\boldsymbol{X}_t$ 为固定值。随着应用问题的复杂化及决策信息的多元化，根据静态信息描述的状态空间很难表征决策环境的全部特征，部分时序特征在优化计算过程中被忽略，往往造成优化不收敛及学习效率降低。改进的指针网络将状态空间分为静态和动态 2 个部分进行构建[13]，即：

$$\boldsymbol{X}_t = \begin{pmatrix} \boldsymbol{s}_t \\ \boldsymbol{d}_t \end{pmatrix} = \begin{pmatrix} s_1, s_2, \ldots, s_M \\ d_{t,1}, d_{t,2}, \ldots, d_{t,M} \end{pmatrix} \tag{11}$$

式中：$\boldsymbol{s}_t$ 表示静态信息；$\boldsymbol{d}_t$ 表示动态信息；M 表示输入的状态信息维数。由于静态动态信息维数相同，可以使用相同的编码及解码器进行特征处理。拓扑优化问题中，由于元素不可重复选取，每个时刻选择的开关编号及状态选择都会直接反映在初始拓扑上。尽管决策方案的验证评价发生在策略构建完毕时，但结合应用实际，为快速减少失电节点，往往需要尽快操作能够恢复最多节点供电的开关。在模型中嵌入表征执行当前元素导致的配网供电节点变化数动态信息，能够准确地表达每步决策对整体方案可靠性的影响。动态信息 d_t 为

$$d_{t,m} = \sum_{i \in D} \gamma_i (G'' = x_m * x_t * \cdots * x_1 * G \,|\, x_t \in \Omega_t) \tag{12}$$

式中：m 代表需要计算动态信息的静态元素位置；x_m 表示其元素操作信息。

需要注意的是，静态动态信息进行编码处理时作为独立变量执行，即其各自分属不同参数的编码器计算。进行解码时，需要对 2 个部分信息作张量合并，共用 1 个解码器进行计算，在注意力机制返回的指针信息中也仅包含静态信息。

2.1.4 掩码设计

由于指针网络的决策依靠注意力机制计算得到的决策概率分布实现，将对应的概率降至 0 可避免元素被选取，此处理方式称为掩码。加入了掩码后的注意力概率计算表达式修正为

$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{h}_t + \log(\boldsymbol{\lambda}_t)) \tag{13}$$

式中：$\boldsymbol{\lambda}_t$ 表示当前时刻 t 的掩码向量，每位的取值均为 0 或 1。当某一位取值为 0 时，对应位置元素被选取的概率计算为 0，不会被选中。掩码的基础功能是控制指针无重复的选取完全集中的元素，即每次预测后将对应元素编号对应的概率置零。利用深度强化学习技术处理优化问题通常需要将约束转化，使智能体动作始终满足约束条件，减少过多约束惩罚对探索式学习的不利影响。在指针网络应用于车辆路径优化等问题求解时，基于约束判断和掩码组合控制备选元素，取得很好的优化效果。

故掩码设计需考虑静态状态信息结构、元素选取逻辑及目标约束转化。静态状态信息结构方面，不同于基础掩码的顺序剔除功能，由于完全集内存在互斥元素，执行预测后掩码不仅需要覆盖已选元素编号，同时也要覆盖其互斥元素编号。由于策略选择的元素不可再次选取，所以此类掩码值置 0 变化不可逆。

元素选取逻辑方面，主要针对序列起始元素的选择是否固定进行讨论。一类组合优化问题需要固定序列起点与终点，此时掩码在第 1 步决策前要覆盖非合法项，在第 2 步决策起取消非法覆盖。而在本模型求解的配网拓扑控制逻辑中，控制组合的起始开关编号不固定，可在第 1 步采用全 1 矩阵的掩码自由决策，此时作为起始解码信号的信息不是静态信息，而是由–1 组成的决策激活信号。

目标约束转化方面，根据式(7)(8)构建的模型约束，可能的策略非法情况多发生于节点电压越限或潮流计算不收敛，可以通过动态状态信息及潮流计算结果进行判别，对于每个非法动作，对相应编号对应的掩码值置 0。此外，式(5)(6)构建的多目标优化问题，可以从掩码设计层面予以简化，通过在固定时间决策轮数屏蔽全部备选元素跳出交互进程，达到控制交互次数及大量无效探索计算的减少，同时避免多目标补偿的效果。

综上所述，以 t 时刻的模型决策为例，掩码的计算逻辑总结如下：

1）根据 t–1 时刻后掩码矩阵，依次将所有非 0 的状态加入决策序列并判断约束条件，将不满足的对应编号掩码值置 0。

2）将 t 时刻选择的状态编号掩码置 0。

3）将 t 时刻选择状态的互斥状态掩码置 0。

4）将 1）置 0 的掩码值还原。

5）若 t 等于限制轮数 T，则掩码矩阵置为全 0。反之，则保存掩码矩阵并用至 t+1 步决策。

2.2　深度强化学习算法

深度强化学习是以交互为核心学习方法，其通过评估由状态信息表征的决策序列的期望奖励值决定策略优劣。其核心要素包括状态空间、动作空间及奖励函数[10]。将强化学习算法应用于组合优化问题不同于传统强化学习决策，后者具有明显的多步决策流程，可以利用累积经验池进行决策更新。以本文的拓扑优化研究为例，奖励评价仅发生在完整决策组合形成之后，对外表现为 1 个单步决策，而内部通过 GRU 单元串联将交互作用以动态信息及隐藏层特征的形式表示，通过奖励函数间接影响隐藏层变量，进而表达决策优劣。本节首先介绍状态空间、动作空间和奖励函数的设置方法。

2.2.1　状态空间设计

深度强化学习关注决策轨迹，故与启发式算法不同，需要更多中间状态及环境状态描述以支撑决策训练。除了式(4)表达的决策元素完全集外，为了将初始拓扑信息及操作数目特征进行表示，在静态状态元素中加入表达开关状态变化的 0-1 变量 β，其在对应编号开关发生变化时取值为 1，反之为 0。以此表达的静态状态空间Ω是从完全集中选择的一半元素组成的集合，可表达为

$$\Omega=\{(N,\beta_N,\alpha_N)\,|\,N=0,1,\ldots,n\} \tag{14}$$

加入 β 变量对元素的操作属性没有影响，仍然用*表示元素操作。由于动作空间没有显式的定义，其本质是对决策元素的选择，在模型中通过注意力机制概率模型进行表达。

2.2.2　奖励函数设计

奖励函数间接表达目标函数值。由于在决策阶段通过掩码方式屏蔽了非法项，不必要设置约束惩罚项。奖励值以判断各节点负荷满足情况为主，以执行拓扑操作统计为辅。构建的组合评价值为

$$R(G',\Omega)=c_1[|D|-\sum_{i\in D}\gamma_i(G')]+c_2\sum_j\beta_j \tag{15}$$

式中：c_1 表示 f_1 可靠性目标评价权重；c_2 表示 f_2 快速性目标评价权重。奖励函数的评价值越低，组合策略方案越满足优化要求，在训练中采用梯度下降的更新方式。由于在本文中构建的单步决策问题中不涉及奖励累计问题，故不必要设置其他参数进行奖励引导。

2.2.3　参数更新算法

本文采用 AC 强化学习算法[10]更新神经网络参数。作为经典的 ACTOR-CRITIC 架构算法，其 ACTOR 网络就是指针网络，其输出并维护 2 个变量：决策元素序列Ω及序列对应的对数概率值集合 Π_p。根据Ω计算组合策略的评价函数值 R，根据 Π_p 计算该组合被选出的概率。采用带基线的策略梯度更新 ACTOR 网络，目的是平抑网络更新方差，其目标函数为

$$J_\theta(\boldsymbol{X})=E_{s\sim p_\theta(\cdot\,|\boldsymbol{X})}R(s|\boldsymbol{X}) \tag{16}$$

计算其策略梯度并进行基线修正，即：

$$\begin{cases}\nabla_\theta J_\theta(\boldsymbol{X})=E_{s\sim p_\theta(\cdot\,|\boldsymbol{X})}[(R(s|\boldsymbol{X})-b_\mu(\boldsymbol{X}))\cdot\\ \qquad\qquad\qquad\nabla_\theta\log p_\theta(s|\boldsymbol{X})]\\ \log p_\theta(s|\boldsymbol{X})=\sum\limits_{p_k\in\Pi_p}\log p_k\\ R(s|\boldsymbol{X})=R(G',\Omega)\end{cases} \tag{17}$$

式中：θ 为指针网络的网络参数；E 代表期望算符；s 代表决策策略；p_θ 代表指针网络根据 $\boldsymbol{X}$ 输出的概率分布；$\boldsymbol{X}$ 为决策状态空间；μ 为 CRITIC 网络参数；b_μ 代表 CRITIC 网络输出。基线的预测由 CRITIC 网络计算得到，根据状态空间完全集对应的静态动态信息解码器输出判断决策难度及预期评价值，预测方法是拟合返回组合策略的评价值，其目标函数为

$$J_\mu(\boldsymbol{X})=E[(R(\pi|\boldsymbol{X})-b_\mu(\boldsymbol{X}))^2] \tag{18}$$

2.2.4　算法流程

模型交互和算法训练以批量样本的方式进行，其中元素选择策略在训练阶段为基于注意力机制概率模型的随机采样，而不是总基于最大概率值选择。训练样本集是由固定拓扑配合故障及开关状态设置衍生的初始拓扑状态集，在训练和测试时从中选取一定数量的拓扑场景进行交互决策。综上所述，实现配电网拓扑控制的强化学习模型训练流程如图 2 所示。

3　实验设计与结果分析

本节主要介绍用于训练的故障状态配网拓扑算例集合构建方法、基于深度强化学习的策略模型训练过程及基于测试结果的验证分析。

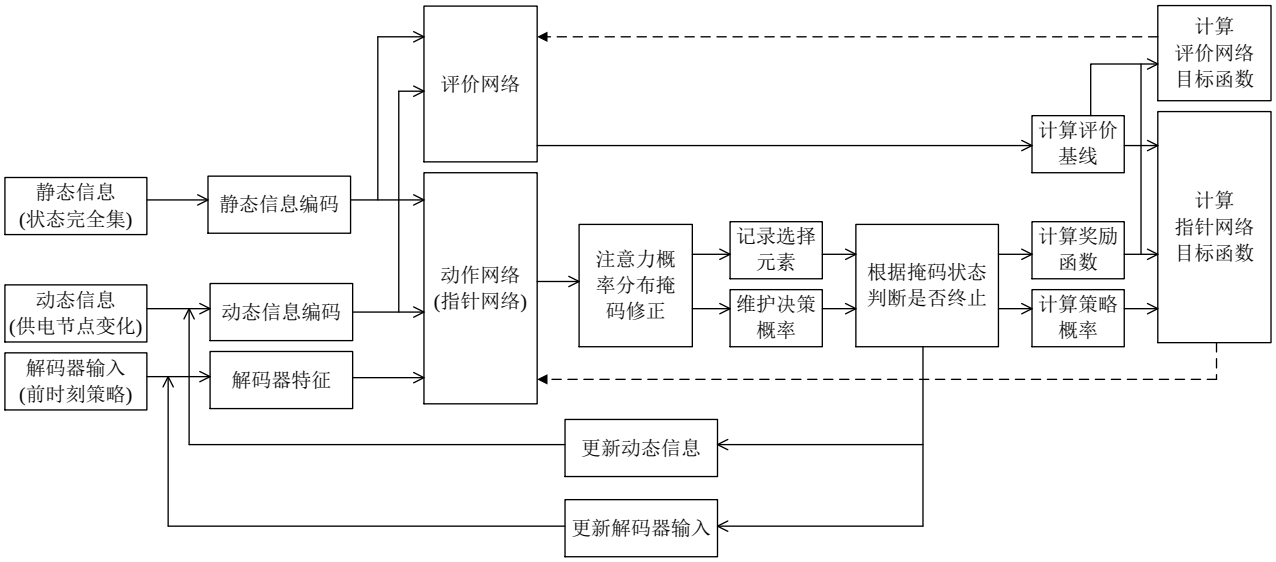

图 2 配电网拓扑控制强化学习模型训练流程

Fig. 2 Training process of reinforcement learning for distribution network topology control

3.1 基础设置

3.1.1 算法软硬件环境

软件方面，核心神经网络搭建及训练主要基于 Pytorch1.4.0 版本完成，配电网仿真部分利用 PandaPower 开源库实现。硬件方面，除必备硬件之外，使用 NVIDIA 1080TI GPU 加速神经网络的训练和预测。

3.1.2 算法运行参数设置

对于 ACTOR 网络的一维卷积编码器，其隐藏层节点数设为 128，解码器采用单层 GRU 单元，隐层节点数设为 128。CRITIC 网络方面，基于已构建的编码器，连接一个输出维度为 1 的 3 层全连接网络实现评价基线输出。2 个网络更新均采用 Adam 优化器进行梯度更新，学习率均设为 10^{-4}。奖励函数方面，c_1 取值为 5，c_2 取值为 1。

3.2 仿真实验设计

本实验采用 IEEE 33 节点配电网模型作为固定的基础拓扑，其编号设置情况如图 3 所示。为满足假设，在 0 号线路设置无限大容量发电机以满足负荷需求。

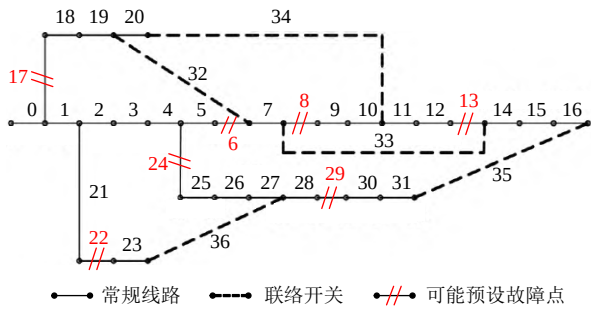

图 3 IEEE 33 节点配电网模型及故障点设置方案

Fig. 3 IEEE 33 distribution network and fault settings

随机假设线路故障发生点极易形成运行孤岛，为简化求解过程中的拓扑状态判断，假定主要故障发生在线路编号为 6、8、13、17、22、24、29 中的 1 个或多个，且至多为 3。原拓扑中的 32、33、34、35、36 为备用联络开关，不考虑开关类型差异，将其视为与常规开关等价。确定故障线路后，除联络开关外随机选取数个设为断开状态。基于以上规则可生成大量不同描述的基础状态空间完全集。实验分为 2 个部分进行，基于单一初始拓扑场景验证算法逻辑有效性，及基于生成样本集测试模型在不同场景下的决策能力。

3.3 实验结果及分析

3.3.1 算法模块功能有效性分析

固定初始拓扑样本集设置为 13、24 号线路发生故障，其他常规开关闭合，联络开关断开。训练样本空间设置为 10，训练批次大小为 10。由于各样本初态相同，意味着同一时刻进行 10 组不同采样进行探索。基于不同设置的算法训练奖励函数变化曲线如图 4 所示。仅静态信息的指针网络训练单

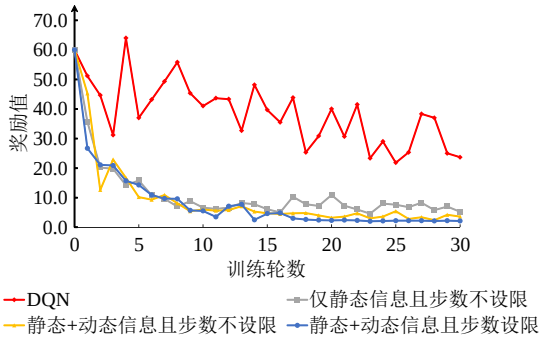

图 4 不同算法的训练奖励变化对比

Fig. 4 Comparison of training rewards of algorithms with different settings

轮约为 12s；加入动态信息但掩码轮数不设限制的单轮训练约 80s；加入动态信息但轮数设限的单轮训练约 32s。模型训练收敛值在 2 附近，表明经过 2 个开关状态变化达到全负荷供电效果，改变的开关编号为 33 和 36，与工程经验及最优策略一致。

对比静态动态信息结合及仅静态信息描述的曲线走势及训练时长可以得出，虽然都可以优化目标值，但仅依靠静态信息无法使模型利用尽量少的开关状态操作达到最优效果，而加入了动态信息描述的状态表征输入能够使模型收敛到更优的决策组合。加入了动态信息作为辅助状态输入后，模型可以用最少的操作完成针对特定故障的恢复送电，验证了模型有效性。

另一方面，对比加入限制掩码更新的固定步长，有效地提升了训练效率，达到稳定收敛的轮数更少。常规掩码需要对全部元素进行一轮概率计算，总计算量明显超过固定步长的方式，使计算时间加长；由于本文设计的拓扑决策方式是基于初始拓扑的变化量决策，即使没有对全部开关状态做判断，亦可以通过当前已选择的状态元素来改变初始拓扑，进而评价决策优劣。另外，减少的动态信息计算即潮流计算量也是效率提升的关键。

3.3.2　算法应用适用性分析

为验证基于指针网络结构的强化学习求解组合优化的适用性，本文对比了使用传统深度强化学习深度 Q 网络方法(deep Q networks，DQN)的拓扑控制训练。使用 DQN 方法需要重新构建适用其算法逻辑的动作 a 表达，利用每个深度神经网络输出层神经元定义对应编号开关的状态，输出神经元激活函数为 tanh 函数，当输出值大于 0 时认为开关闭合，反之断开。以此方式完成 1 个维度与开关数目相同的动作空间构建。对比 DQN 和本文方法可以看出，基于 DQN 的训练虽然有下降趋势，但很难在短时间内发现有效的控制策略，其中动作维度过大以及状态空间的信息稀疏是主要原因，考虑到问题属于单步决策，亦使得适合多步决策的状态动作值函数失去意义。

3.3.3　混合样本训练下的算法有效性分析

通过 3.2 节描述的方式生成包含 50 个样本的初始状态集进行训练。经试验统计，一轮训练约耗时 210s。模型在训练过程中，测试分数会在一段轮数内保持稳定，在某些轮数发生下降。混合样本训练的模型能力测试曲线如图 5 所示，由图可见，约在 100 轮附近模型达到测试收敛。经验证明，完成训练的决策模型同样可以实现在大部分故障下的端到端开关控制组合策略输出，收敛值在 4 附近表明在整个样本集内本文模型策略能够大概率避免失电负荷。由于训练样本规模及超参数调节限制，很难完全调整模型收敛至理论上的最优参数结构，故很难让每个状态下都找到精确最优收益。由于多数情况联络开关在拓扑内承担应急转供的角色，作为故障处理的优先控制对象，即使部分情况下联络开关不必要动作，策略模型也会因探索不完全带来的期望偏差选择其状态改变，进而带来不必要惩罚。但从整体来看，基于指针网络及强化学习实现的拓扑控制决策参数模型实现了研究预期目标。

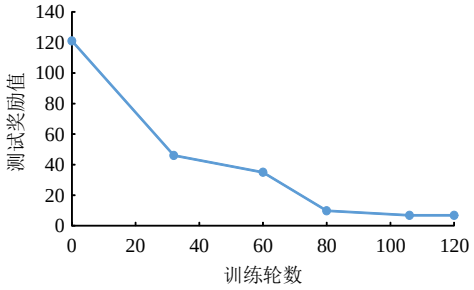

图 5　混合样本训练的模型能力测试曲线

Fig. 5　Model ability test based on mixed sample

4　结论与展望

本文将基于指针网络的深度强化学习组合优化技术应用于 IEEE 33 配电网的故障供电恢复策略研究中，主要结论如下：

1）设计适用于深度学习模型的有效配电网拓扑状态表征和决策动作集。

2）创新地设计掩码机制转化优化问题复杂度并提升训练效率。

3）实现适用于多类故障的神经网络自学习和端到端控制策略计算模型训练方法。经验证，在预设条线路上随机设置故障组合，可通过神经网络直接计算对应的核心联络开关控制方式，满足研究问题假设与预期。

本研究转化了配电网拓扑优化与强化学习决策的矛盾，为该技术路线在更复杂场景上的研究应用提供参考。

参考文献

[1] 马钊，刘颖异，尚宇炜，等．CIGRE 2016 未来电力系统及主动配电系统技术新动向[J]．中国电机工程学报，2017，37(1)：27-36．
MA Zhao，LIU Yingyi，SHANG Yuwei，et al．CIGRE 2016 development trends of future power system and active distribution system[J]. Proceedings of the CSEE，2017，37(1)：27-36(in Chinese)．

[2] 李旸，唱友义，梁晓赫．新能源接入的配电网效益最优重构方法研究[J]．控制工程，2021，28(5)：931-937．
LI Yang，CHANG Youyi，LIANG Xiaohe．Research on an optimal reconfiguration method of distribution network benefits with new

energy access[J]. Control Engineering of China，2021，28(5)：931-937(in Chinese).

[3] JUFRI F H，WIDIPUTRA V，JUNG J. State-of-the-art review on power grid resilience to extreme weather events：definitions，frameworks，quantitative assessment methodologies，and enhancement strategies[J]. Applied Energy，2019(239)：1049-1065.

[4] 祝旭焕，童宁，林湘宁，等. 基于柔性多状态开关的主动配电网负荷在线紧急转供策略[J]. 电力系统自动化，2019，43(24)：87-95.
ZHU Xuhuan，TONG Ning，LIN Xiangning，et al. Flexible multi-state switch based online emergency load transferring strategy for active distribution network[J]. Automation of Electric Power Systems，2019，43(24)：87-95(in Chinese).

[5] 刘永梅，王金丽，杨红磊，等. 计及柔性负荷调节能力的有源配电网动态优化方法[J]. 高电压技术，2021，47(1)：73-80.
LIU Yongmei，WANG Jinli，YANG Honglei，et al. Dynamic optimal method of distribution network in consideration of flexible load adjustment capability[J]. High Voltage Engineering，2021，47(1)：73-80(in Chinese).

[6] 邓斯凯，毛弋. 基于量子人工蜂群算法的配电网多目标优化重构[J]. 湖南师范大学自然科学学报，2021，44(2)：80-86.
DENG Sikai，MAO Yi. Multi-objective optimal reconfiguration of distribution network based on quantum artificial bee colony algorithm[J]. Journal of Natural Science of Hunan Normal University，2021，44(2)：80-86(in Chinese).

[7] 甄霖腾. 计及新能源消纳的主动配电网多目标优化调度策略研究[D]. 保定：河北农业大学，2021.

[8] MAROT A，DONNOT B，ROMERO C，et al. Learning to run a power network challenge for training topology controllers[J]. Electric Power Systems Research，2020(189)：106635.

[9] YOON D，HONG S，LEE B J，et al. Winning the L2RPN challenge：power grid management via semi-markov afterstate actor-critic[C]// Proceedings of the 9th International Conference on Learning Representations. Austria：OpenReview.net，2021.

[10] 刘朝阳，穆朝絮，孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报，2020，2(4)：314-326.
LIU Zhaoyang，MU Chaoxu，SUN Changyin. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology，2020，2(4)：314-326(in Chinese).

[11] 李凯文，张涛，王锐，等. 基于深度强化学习的组合优化研究进展[J]. 自动化学报，2021，47(11)：2521-2537.
LI Kaiwen，ZHANG Tao，WANG Rui，et al. Research reviews of combinatorial optimization methods based on deep reinforcement learning[J]. Acta Automatica Sinica，2021，47(11)：2521-2537(in Chinese).

[12] BELLO I，PHAM H，LE Q V，et al. Neural combinatorial optimization with reinforcement learning[C]//Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon：Université de Montreal，2017.

[13] NAZARI M，OROOJLOOY A，OROOJLOOY L V，et al. Reinforcement learning for solving the vehicle routing problem[C]//Advances in Neural Information Processing Systems 31：Annual Conference on Neural Information Processing Systems 2018. Montréal：Curran Associates Inc，2018：9861-9871.

[14] 胡尚民，沈惠璋. 基于强化学习的电动车路径优化研究[J]. 计算机应用研究，2020，37(11)：3232-3235.
HU Shangmin，SHEN Huizhang. Research on electric vehicle routing problem based on reinforcement learning[J]. Application Research of Computers，2020，37(11)：3232-3235(in Chinese).

[15] VINYALS O，FORTUNATO M，JAITLY N. Pointer networks[C]//Advances in Neural Information Processing Systems 28：Annual Conference on Neural Information Processing Systems 2015. Quebec：Curran Associates Inc，2015：2692-2700.

[16] 刘帅，孔亮，刘自发，等. 基于深度强化学习的输电网网架规划方法[J]. 电力建设，2021，42(7)：101-109.
LIU Shuai，KONG Liang，LIU Zifa，et al. Transmission network planning method based on deep reinforcement learning[J]. Electric Power Construction，2021，42(7)：101-109(in Chinese).

[17] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach：Curran Associates Inc，2017.

[18] CHO K，VAN MERRIENBOER B，GÜLÇEHRE Ç，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha：SIGDAT，2014.

闫冬

在线出版日期：2021-12-01。
收稿日期：2021-07-14。

作者简介：
闫冬(1993)，男，工程师，通信作者，研究方向为深度强化学习等人工智能技术在电力系统中的应用，E-mail：yandong@epri.sgcc.com.cn；
彭国政(1945)，男，高级工程师，研究方向为人工智能技术在电力系统中的应用，E-mail：pengguozheng@epri.sgcc.com.cn。

（责任编辑　温杰）