文章编号:1007-5321(2020)02-0087-07 **DOI**:10.13190/j.jbupt.2019-103

基于深度强化学习的综合能源业务通道优化机制

马庆刘[1], 喻 鹏[1], 吴佳慧[1], 熊 翱[1], 颜 拥[2]
(1. 北京邮电大学 网络与交换技术国家重点实验室, 北京 100876;
2. 国网浙江省电力有限公司, 杭州 310007)

摘要: 为了保障综合能源系统的稳定运行,承载综合能源业务的通信网络需要具备高可靠、低风险等特征. 依据综合能源业务的通道要求,提出了一种深度强化学习的算法,旨在对大规模综合能源业务在承载的电力通信网上寻找到整体最优的路径. 该方法以整体时延和网络负载均衡度为目标,对网络拓扑进行训练,并保存模型,然后通过迭代学习获取最优的结果. 仿真结果表明,该方法找到的路径既可以保证整体时延较短,又可以保证网络的整体负载均衡. 同时,在网络规模很大、业务数量很多的情况下,深度强化学习算法可有效提高计算效率.

关 键 词: 深度强化学习; 路径优化; 时延; 负载均衡
中图分类号: TN929.11 **文献标志码**: A

A Integrated Energy Service Channel Optimization Mechanism Based on Deep Reinforcement Learning

MA Qing-liu[1], YU Peng[1], WU Jia-hui[1], XIONG Ao[1], YAN Yong[2]
(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. State Grid Zhejiang Electric Power Company Limited, Hangzhou 310007, China)

Abstract: In order to ensure the stable operation of the integrated energy system, the integrated energy service needs to have high reliability and low risk when being carried by the communication network. According to the channel requirements of the integrated energy service, an algorithm of deep reinforcement learning is proposed, aiming to find the overall optimal path for the large-scale integrated energy service on the carried power communication network. The method that aims at the overall delay and network load balance, trains the network topology and saves the model, and then obtains the optimal result through iterative learning. The simulation results show that the routing found by this method can ensure the overall delay is short and guarantee the overall load balance of the network. At the same time, for scenarios with a large network size and a large number of services, the deep reinforcement learning algorithm can effectively improve the computational efficiency.

Key words: deep reinforcement learning; routing optimization; time delay; load balancing

综合能源系统是一种新型能源供应系统,它使 用各种能源和新型能源,满足设计领域的电力、热力

收稿日期: 2019-05-31
基金项目: 国家电网公司科技项目"高可信智能感知互动综合服务系统关键技术研发及应用示范"(52110418002V)
作者简介: 马庆刘(1994—), 男, 硕士生.
通信作者: 喻 鹏(1986—), 男, 副教授, E-mail: yupeng@bupt.edu.cn.

和冷却负荷的综合能源需求[1-2]. 综合能源系统以电能为核心,通过各种电加热和电制冷设备将电能转换为所需的能源形式. 它满足该区域非采暖/空调期间的电力负荷需求、采暖期间的热负荷需求和空调期间的冷负荷需求. 可设计冷藏和蓄热系统以实现能量的分时利用,同时,可设计分布式发电单元和热泵系统,以充分利用可再生能源和地热能. 在综合能源系统的规划、建设和运营过程中,经过有机协调和优化能源生产,输配电(能源网络),转换,储存和消费,形成能源生产、供应和营销的综合系统. 综合能源系统主要由能源供应网络和能源交换链路组成,将储能链路、终端集成能源供应单元以及大量终端用户组合在一起. 综合能源业务则是指综合能源系统提供服务过程中端到端的通信过程. 图1所示为综合能源系统和对应通信网络的示例.

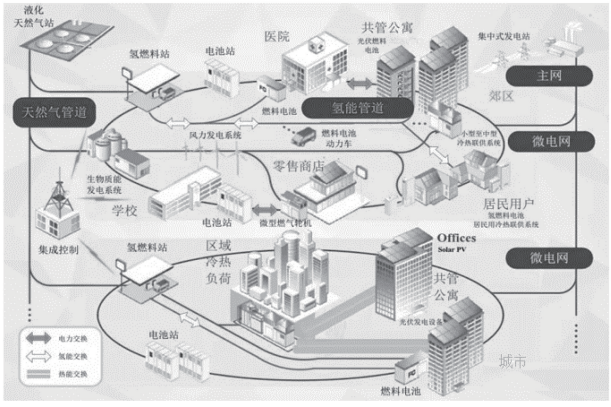

图1 综合能源系统

为了保证综合能源系统的稳定可靠运行,有必要在通信网络上进行综合能源业务进行规划和部署时,对业务路由进行优化,以实现多业务整体的高效和低风险运行. 在电力通信网的相关路由优化机制上,一些研究成果[3]以业务QoS和整体负载均衡为优化目标,提出了改进QoS的路由配置算法,有效地提升了业务路由的QoS水平和负载均衡安全性. 然而这些方法的效率往往较低,他们很难在有效时间范围内,在兼顾整体效率和负载均衡[4]方面,找到最优的路由. 同时,这些业务尚未考虑综合能源业务的需求,也未考虑深度强化学习等新型算法. 因此,利用了深度强化学习算法,将其用于业务通道的优化,既可保障业务的总体QoS,又可兼顾网络的负载均衡,同时可以弥补传统方法的低效,有效地实现业务通道的优化.

1 网络系统模型

1.1 网络模型

为了抽象综合能源系统中的业务通道优化问题,首先对综合能源系统的通信网的网络拓扑进行建模,将通信网定义成为一个有权无向网络,并有如下假设:

1) 所有节点的处理时延看成相同,每两个节点之间的传输时延作为网络权重,节点可以是医院、零售站、学校等;

2) 任意站点之间的链路作为全网的边,系统通信方式为双向通信,因此认为边均为有权无向边.

通信网可用节点数为 N,链路数为 M 的网络拓扑图 $G(V,E)$ 表示,其中 $V=\{v_1,v_2,\cdots,v_n\}$ 代表节点的集合,$E=\{e_1,e_2,\cdots,e_m\}$ 代表边的集合,边的权重由该条边上的传输时延决定,时延的计算为[5]

$$t_1 = Ln_1/c \tag{1}$$

其中:L 为光缆长度;n_1 为光纤折射率,对于G.652光缆而言,n_1 的值为1.48;c 为光速,$c=3\times10^5$ km/s. 所以,单位长度光缆的时延约为5 μs/km,即 $t_1=5L$.

假设有 L 个业务,对每个业务,两点之间的路径表示为 p_{ij},其中起点和终点分别为 v_i 和 v_j. 为了方便后续进行多业务的优化,每个业务两点间的路径采用深度优先遍历的方法[6],寻找到所有的满足约束条件的无环路径,用矩阵 $\boldsymbol{C}_k$ 表示

$$\boldsymbol{C}_k=\begin{pmatrix} v_i & \cdots & v_j & 0 & \cdots & 0 \\ v_i & \cdots & \cdots & v_j & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_i & \cdots & v_j & 0 & \cdots & 0 \end{pmatrix} \tag{2}$$

其中 $\boldsymbol{C}_k$ 表示第 l 个业务的所有路径的矩阵表示,矩阵行数表示无环路径条数,矩阵列数为整个拓扑图的节点个数 N,矩阵每行的 v_i 到 v_j 之间表示每条路径的节点序列,后面的0作为填充. 对每条链路计算时延 t,可用向量 $\boldsymbol{T}_k=\{t_1,t_2,\cdots,t_l\}$ 表示第 k 条业务对应的所有路径的时延.

1.2 目标函数

由于通信网本身结构在实际应用场景中是很难被改变的,所以为了降低链路中断所造成的影响,从负载均衡的角度,对电力通信网传输网上承载的业务路由进行优化,提出相应的优化策略,以降低链路中断所带来的影响. 因此,考虑了对业务通信时长

和全网业务均衡度的影响，以各业务的传输延时和全网业务负载均衡度的加权和最小化为优化策略的目标函数. 为了使业务能够均匀的分布在网络上，每条链路均匀的承载业务，需要对各链路的业务容量进行一定的限制. 则智能电网通信传输网络路由优化策略的数学模型为

$$\min(\alpha\overline{T}+\beta D)$$

$$\overline{T}=\sum_{i=1}^{L}T_i/L$$

$$D=\sqrt{\sum_{i=1}^{m}\left(L_i-\sum_{i=1}^{m}L_i/m\right)^2\Big/m} \tag{3}$$

其中，$\overline{T}$ 为整个网络所有业务的平均传输时长，T_i 为业务 i 的传输时延，L 为总的业务个数，D 为全网所有链路的均衡度，L_i 为第 i 条链路上所承载的业务个数，m 为全通信网链路总个数，$T_{\max}$ 为所允许的最大业务时延，$L_{\max}$ 为所允许的链路承载的最多业务个数. α 和 β 为 2 个常系数. $T_i<T_{\max}$，$L_i<L_{\max}$ 分别表示时延最大约束和任意链路业务数最大约束.

该目标是业务时延和均衡度的加权和，根据 α 和 β 的值对业务时延和均衡度的重要度进行加权，最终目标是让它们的加权和最小，以达到整体最优. 为了求此最优解，在第 3 节中采用了深度强化学习算法，对模型进行优化求解，由于多业务路径优化问题的状态空间非常大，因此使用接近遍历的方法来求解是不可取的. 深度强化学习算法只对部分状态数据进行训练，便可得到较为理想的结果，因此对于该问题而言，此方法既可以很好地逼近最优解，又可以大幅提高计算效率.

2　基于深度强化学习的路由优化策略

2.1　深度强化学习算法

Q-learning 算法和 $Q(\lambda)$ 算法[7] 这 2 种无监督学习方法虽然都可以自己从周围的环境中进行学习，然而，他们仍然需要人工设计相应的特征以便训练收敛的 Q 矩阵. 在实际场景中，状态的数量往往是非常大的，而且在许多场景下，特征也是难以用人工设定的. 而神经网络正好是针对大量数据所设计而成的，所以在这里考虑用神经网络替换 Q-learning 和 $Q(\lambda)$ 中的 Q 值表[8]. 这就形成了现在所谓的 Deep Q-Learning 算法.

Deep Q-Learning 算法模型结合了强化学习模型和神经网络模型，任意状态 s，对应输出一个动作的向量 $\boldsymbol{Q}(s,*;\theta)$，其中 θ 为神经网络的参数. n 维的状态空间对应的动作空间是 m 个动作，因此神经网络本质是 n 维状态空间到 m 维动作空间映射的函数. Deep Q-Learning 算法有 2 个核心[9]：目标网络、经验回放. 目标网络的计算更新公式为

$$Q(s,a)=$$
$$Q(s,q)+\alpha(r+\gamma\max_a,Q(s',a')-Q(s,a)) \tag{4}$$

对于强化学习中的经验回放，首先存储观察到的状态转换. 在样本累积到一定程度后，从中随机采样以更新网络. 经验回放是深度强化学习的一个非常重要的部分，它极大地提高了深层强化学习的系统性能.

基于 Deep Q-Learning 多业务路由规划算法，有两部分非常重要：其一，计算 CNN 神经网络中的损失函数，并利用梯度下降法训练网络参数；其二，在训练网络参数的过程中提取训练样本.

1）损失函数的计算

Q-Learning 中的旧 Q 值与目标 Q 值之间的关系，对应 Deep Q-Learning 中神经网络训练过程中的结果值和输出值之间的关系. 因此，定义损失函数的为

$$L(\theta)=E[(Q'-Q(s,a;\theta))^2] \tag{5}$$

$$Q'=r+r\max_a,Q(s',a') \tag{6}$$

2）训练样本提取

基于 Deep Q-Learning 路由规划算法的经验回放的概念是：不同的业务随机探索网络路径，并将学习的经验放入记忆池中. 当经验累积到一定程度时，从记忆池中随机选取部分样本进行训练. 采用随机抽样的原因是：不同业务随机学习周围环境而获得的样本是与时间具有相关性的. 由于时间相关性，如果数据直接用作训练和更新 Q 值表的样本，则系统收敛将受到很大影响. 因此，采用随机抽样的方法来解决时间相关问题.

2.2　算法实现

针对上述建立的电力通信网路由优化模型，采用深度强化学习的方法，尽可能从网络中为各业务寻求总体时延最短的路径，同时来保证网络的业务均衡度.

该算法的学习实体为所有的业务集合，其状态空间为

$$\boldsymbol{S}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_l\} \tag{7}$$

其中：l 为全网业务数，$\boldsymbol{p}_i(i=1,2,\cdots,l)$ 表示第 i 个业务的一条路径，即矩阵 $\boldsymbol{C}_k$ 中的一行，这里 $\boldsymbol{p}_i(i=1,2,\cdots,l)$ 是一个维数为 N 的向量，$\boldsymbol{p}_i(i=1,2,\cdots,l)$ 的取值空间即为矩阵 $\boldsymbol{C}_k$ 中的所有行向量，用 P_i 表示矩阵 $\boldsymbol{C}_k$ 中所有行向量的集合，$\boldsymbol{p}_i\in P_i(i=1,2,\cdots,l)$.

为了更好地定义动作空间，先将每条业务的路径空间 $P_i(i=1,2,\cdots,l)$ 中的所有路径按行循环排序，对每个业务定义动作空间为

$$a_i\in\{-1,0,1\} \tag{8}$$

其中：-1 表示将 $\boldsymbol{p}_i(i=1,2,\cdots,l)$ 向上移动一行，1 表示将 $\boldsymbol{p}_i(i=1,2,\cdots,l)$ 向下移动一行，0 表示 $\boldsymbol{p}_i(i=1,2,\cdots,l)$ 不变. 这样，整个动作空间的大小为 3^l，即对于每个状态 S，都有 3^l 个可选动作可供选择. 整个动作空间为

$$\boldsymbol{A}=\{a_1,a_2,\cdots,a_l\} \tag{9}$$

深度强化学习包括离线构造网络阶段和在线深度 Q 学习阶段 2 个阶段. 离线阶段使用 CNN 获得状态-动作对(s,a)和值函数 $Q(s,a)$之间的关系，值函数是在状态 s 下执行动作 a 时的累积折扣奖励. 在线学习过程中，在每个时段，深度强化学习利用 CNN 得到估计 Q 值，用贪心的方式以 ε 的概率随机选择一个动作 a，而以 $1-\varepsilon$ 的概率选择那个 Q 值最大的动作. 在与环境交互中观察到立即奖励 r 和下一状态 s'，将状态转换(s,a,r,s')存入记忆池中，最后从记忆池中抽样训练更新 CNN 的参数.

因此，根据第 2 节对目标函数的分析，定义立即奖励为

$$r=\begin{cases}\dfrac{1}{\alpha\overline{T}+\beta D}, & \text{满足约束}\\ 0, & \text{其他}\end{cases} \tag{10}$$

目标函数是求 $\alpha\overline{T}+\beta D$ 的最小值，因此当 $\alpha\overline{T}+\beta D$ 越小，就给出越大的奖励，这里对 $\alpha\overline{T}+\beta D$ 取倒数作为立即奖励. 对于不满足约束条件式(3)的，将其立即奖励定义为 0.

2.3 算法流程

基于深度强化学习的业务通道优化算法步骤如下：

1）初始化全网的状态 S，初始化内存池，并设置一个观察值，即变化的最大步数；

2）在当前状态 S 的基础上，选择动作 A，获取相应的奖励值 R，动作结束后的状态 S'，并将相关参数 S,A,R,S'保存到记忆池中；

3）判断记忆池中存储的数据量是否超过观察值，如果不够，转到 4），如果数据足够，转到 5）；

4）判断是否达到之前设置最大查找步数

① 若达到最大查找步数，给 S 随机重置一个状态；

② 若查找未达到最大步数，将当前状态 S 更新为 S’；

返回步骤 2）；

5）开始训练

① 从内存池中随机选取一部分数据作为训练样本；

② 将随机抽样的状态 S'作为训练样本，得到相应状态的 Q 值表；

③ 根据公式计算与 Q 值表对应的 target$Q$ 值；公式为 $Q(S,A)=R+\gamma\max[Q(s',\text{all\_actions})]$

6）使用 Q 值表与 target$Q$ 值来训练神经网络；

7）结束.

基于深度强化学习的业务通道优化算法流程如图 2 所示.

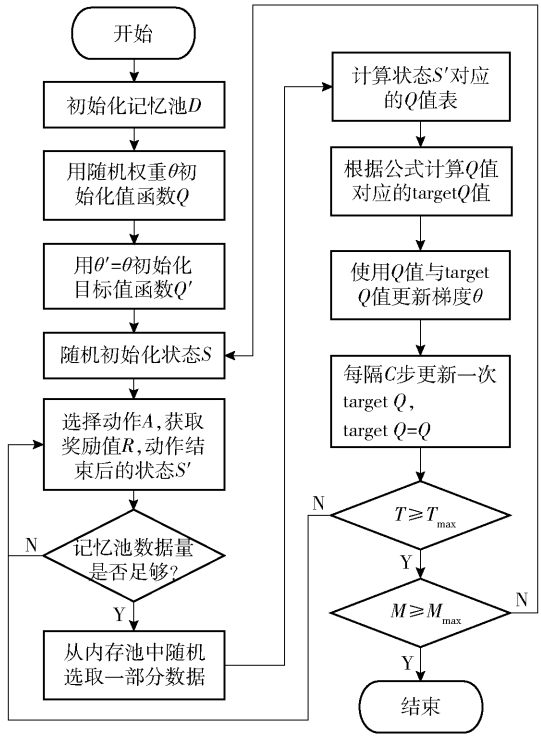

图 2 基于深度强化学习的业务通道优化算法流程

3　仿真与分析

对某地区电力通信网拓扑进行路由优化及分析，拓扑图如图 3 所示，总共设有 14 个路由节点，节点之间的权重值表示传输时延代价. 现有 5 条源—目的业务：2—13，1—9，3—14，5—8，3—6. 每条业务的备选路径在 20 个左右，状态的量级是 20^5，目标是为这 5 条业务分别从备选路径中寻求路径，使得目标函数 $\min(\alpha \overline{T}+\beta D)$ 最小.

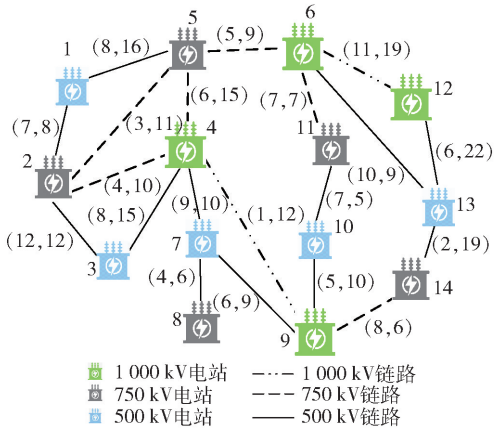

图 3　某地区电力通信网拓扑图

算法中的参数实际取值如表 1 所示.

表 1　算法中使用的参数值

参数	值
目标函数中参数 α,β	0.1,0.9
每条链路最大承载业务数 L_{max}	2
最大业务时延 T_{max}/ms	30
学习率 $\propto$	0.000 1
折扣因子 γ	0.9
贪婪值 ε	0.01 ~ 1
记忆库更新迭代次数 τ	100
观察步数 s	200
记忆库单元大小 D	500
训练回合数 e	2 100

经过学习之后，对模型进行保存，然后再进行测试，网络会寻找到相对理想的路径，对于这 5 条业务而言，算法找到的最优路径见表 2. 其中表格最后一列表示时延代价，括号里面表示最小时延代价. 可以看出，各自的时延并不一定是最小的，但是综合平均时延和网络负载均衡度可以使目标函数达到理想值，整体收益尽可能好.

表 2　深度强化学习算法各业务对应路径

业务	源—目的	对应路径	实际代价
1	2—13	2—5—6—13	18(15)
2	1—9	1—2—4—9	12(12)
3	3—14	3—4—9—14	17(17)
4	5—8	5—4—7—8	19(17)
5	3—6	3—2—5—6	20(19)

业务 1 算法路径选择与最小路径对比如图 4 所示，上方实线箭头表示算法求得的业务 1 的路径选择，下方虚线箭头表示业务 1 时延代价最小的路径，没有选择下方虚线路径的原因是，综合考虑另外几条业务，下方虚线路径所选的链路整体负载较大，会降低网络安全性，因此算法从整体角度考虑，为业务 1 选择上方粗线路径.

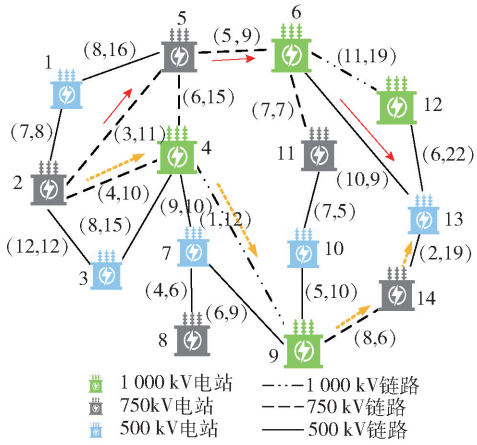

图 4　业务 1 算法路径选择与最小路径对比

算法训练完成之后，对测试结果，每隔 100 步，对业务均衡度、平均时延已经目标函数值进行统计，如图 5 ~ 7. 可以发现，在 1 100 步之前，三者的下降速率很快，达到 1 100 步之后，算法接近收敛，目标函数的曲线基本趋于平稳，此时的解基本接近预期的最优，在 2 100 步左右停止算法迭代，分别得到当前时刻 5 条业务的路径，视为最优路径（见表 2）.

图 8 所示为几种不同算法的性能对比. 这里横轴状态规模表示状态数量，从 1 万个状态到 20 万个状态. 纵轴表示目标函数 $\alpha \overline{T}+\beta D$ 首次达到 2.17 时算法所用的时间（DQN 算法只统计训练时间）.

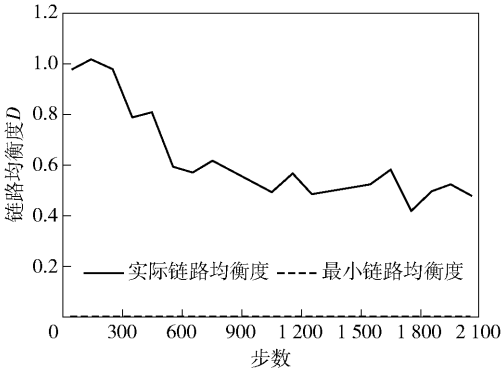

图5 算法迭代过程中全网业务均衡度变化

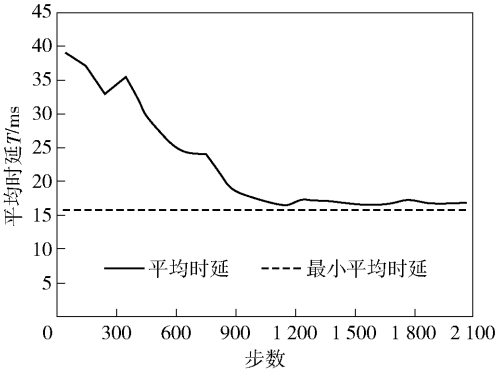

图6 算法迭代过程中全网业务平均时延变化

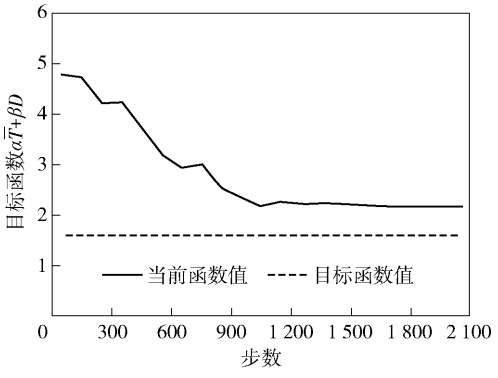

图7 算法迭代过程中目标函数值变化

可以看出,在状态数较少时,深度强化学习并不占优,这是因为深度强化需要一定的数据规模和训练量支撑. 而另外3种算法是接近遍历的方式求解,在状态量少的时候,选择这3种方式是可取的. 当状态数逐渐变大的时候,想要求得最优解,这3种算法的时间复杂度几乎以线性的方式增长,而深度强化学习算法逐渐趋于平缓,这是因为该算法只需要部分数据进行训练,得出模型,便可很好地预测大部分状态对应的情况.

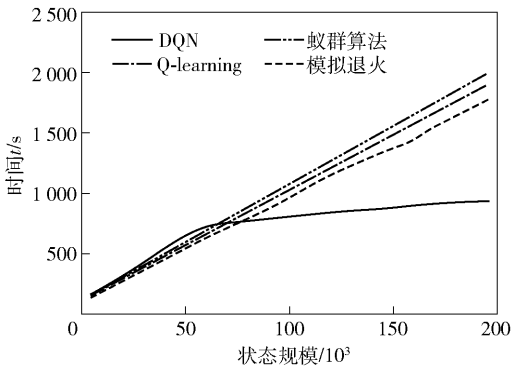

图8 各算法性能对比

4 结束语

针对综合能源业务的应用场景,提出了深度强化学习算法,并将其应用在通道优化机制中,寻求最优路径. 仿真结果表明,在对于状态量很多的该问题时,在较短的时间内即可寻找到相对理想的路径,既可以保证整体时延,又可以保证网络的整体负载均衡,是一种高效实用的方法. 在未来综合能源业务大规模网络中,可以快速地为多业务进行路径选择与优化.

参考文献:

[1] 余晓丹, 徐宪东, 陈硕翼, 等. 综合能源系统与能源互联网简述[J]. 电工技术学报, 2016, 31(1): 1-13.
Yu Xiaodan, Xu Xiandong, Chen Shuoyi, et al. A brief review to integrated energy system and energy Internet [J]. Transactions of China Electrotechnical Society, 2016, 31(1): 1-13.

[2] 贾宏杰, 王丹, 徐宪东, 等. 区域综合能源系统若干问题研究[J]. 电力系统自动化, 2015, 39(7): 198-207.
Jia Hongjie, Wang Dan, Xu Xiandong, et al. Research on some key problems related to integrated energy systems [J]. Automation of Electric Power Systems, 2015, 39 (7): 198-207.

[3] 蔡伟, 杨洪, 熊飞, 等. 考虑电力通信网可靠性的业务路由优化分配方法[J]. 电网技术, 2013, 37(12): 3541-3545.
Cai Wei, Yang Hong, Xiong Fei, et al. An optimized service routing allocation method for electric power communication network considering reliability [J]. Power System Technology, 2013, 37(12): 3541-3545.

[4] 王浩, 李知航, 潘志文, 等. LTE网络中具备QoS保障的动态负载均衡算法[J]. 中国科学: 信息科学,

2012，42(6)：674-686.
Wang Hao，Li Zhihang，Pan Zhiwen，et al. QoS guaranteed dynamic load balancing algorithm in 3GPP LTE networks[J]. Scientia Sinica (Informationis)，2012，42(6)：674-686.

[5] 高钧利. SDH 光传输系统的时延测算[J]. 浙江电力，2011，30(4)：42-45.
Gao Junli. Time delay test and calculation in SDH，based optical transmission system[J]. Zhejiang Electric Power，2011，30(4)：42-45.

[6] 周泰. 图的深度优先遍历算法及运用[J]. 电脑编程技巧与维护，2011(16)：93-94.
Zhou Tai. The DFS for graph and its application[J]. Computer Programming Skills & Maintenance，2011(16)：93-94.

[7] 高阳，陈世福，陆鑫. 强化学习研究综述[J]. 自动化学报，2004，30(1)：86-100.
Gao Yang，Chen Shifu，Lu Xin. Research on reinforcement learning technology：a review[J]. Acta Automatica Sinica，2004，30(1)：86-100.

[8] Mnih V，Kavukcuoglu K，Silver D，et al. Human，level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[9] 刘建伟，高峰，罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报，2019，42(06)：1406-1438.
Liu Jianwei，Gao Feng，Luo Xionglin. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers，2019，42(06)：1406-1438.