与监督学习、无监督学习等不同，强化学习解决的是另一类的动态决策问题。强化学习最早的研究是为了解决控制问题，上世纪60年代，贝尔曼在研究最优控制理论过程中，提出了著名的贝尔曼最优方程[1]，奠定了强化学习的理论基础。1988年，Sutton首先提出了时间差分算法[2]，可以更高效的利用与环境的交互数据，采用自举的方式无模型地实现强化学习问题的求解。1989年，另一种以值函数为基础的无模型Q学习算法也很快面世[3]。20世纪初，Sutton又提出了策略梯度法，可以直接对策略进行学习[4]。至此，强化学习的理论框架基本搭建完毕。

强化学习的理论发展很快在不同领域中得到应用，然而传统的强化学习算法受限于维度灾难问题，很难应用于状态空间与动作空间很大的问题。随着深度学习的快速发展将深度学习与强化学习相结合的深度强化学习方法开始出现，2013年Deep Mind公司提出了DQN模型[5]，用深度卷积神经网络解决状态空间爆炸的问题，并在电子游戏上达到甚至超过了人类水平。2015年，Deep Mind公司又提出了基于深度强化学习的围棋算法AlphaGo[6]，打败了世界围棋冠军李世石，使得深度强化学习开始走进大众的视野。

深度强化学习方法主要可以分为三种——基于值函数的方法、基于策略的方法和行动者-评论家方法。

基于值函数的方法以当前观测到的环境状态作为模型输入，以动作价值函数作为模型输出。其代表就是DQN，DQN直接以游戏画面的原始像素作为模型的输入，然后用深度卷积神经网络抽取其中的关键特征，并拟合得到动作空间的动作价值函数，得到最优控制策略。此后，针对DQN存在的过优化等问题，又有DDQN[7]、Dueling DQN[8]等算法相继问世，在大多数的游戏上，基本都达到甚至超过了人类水平。

基于策略的方法是将传统的策略梯度法与深度神经网络相结合，直接用深度神经网络拟合策略，以当前观测到的环境状态作为模型输入，以动作或动作的概率分布作为模型输出。基于策略的方法直接以动作作为输出，无需进行离散化出力，因此可以有效解决动作空间的维度爆炸问题，实现连续动作的控制。

行动者-评论家方法（Actor Critic，AC）是将基于值函数的方法与基于策略的方法相结合，因此具备两种方法的优点，是目前研究的热点方向。优势函数行动者评论家算法（A2C）用动作优势函数代替AC算法中的累积回报，大大提高了模型的收敛性。为了提高数据的利用效率，PPO算法采用重要性采用的思想，提出了异策略的A2C算法[9]。为了解决数据的相关性问题，Deep Mind采用多个仿真环境并行的方式提出了异步优势行动者评论家算法（A3C）[10]，在大多数任务上达到了目前最好的效果。除此之外，针对确定性的动作策略，也有一些算法根据AC框架被相继提出，如DPG[11]和DDPG算法[12]等。

传统的强化学习方法在电压控制与无功优化问题中已有相关研究[13-15]，但是传统强化学习算法通常难以有效解决状态空间与动作空间的维度灾难问题，在实际的大规模系统中难以有效应用。同时目前已有研究主要还是基于模型依赖的强化学习算法，难以适用于低感知度配电网。将深度强化学习用于电力系统，用数据驱动无模型地解决动态优化问题的目前相关研究开展仍处于初步阶段，相关文献较少。

**参考文献**

[1] Bellman R. On the theory of dynamic programming. Proceedings of the National Academy of Sciences of the United States of America, 1952, 38(8): 716

[2] Sutton R S. Learning to predict by the methods of temporal differences. Machine learning, 1988, 3(1): 9～44

[3] Watkins C J C H. Learning from delayed rewards, 1989

[4] Sutton R S, Barto A G. Reinforcement learning: An introduction MIT press, 2018

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

[6] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. nature, 2016, 529(7587): 484

[7] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning, 2016

[8] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning. arXiv preprint arXiv:1511.06581, 2015

[9] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017

[10] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning, 2016, 1928～1937

[11] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms, 2014

[12] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015

[13] 陈星莺, 钱锋, 杨素琴. 模糊动态规划法在配电网无功优化控制中的应用. 电网技术, 2003, (02): 68～71

[14] Hsu Y Y, Kuo H C. Dispatch of capacitors on distribution system using dynamic programming. Generation Transmission & Distribution Iee Proceedings C, 2002, 140(6): 433～438

[15] Lu F C, Hsu Y Y. Fuzzy Dynamic Programming Approach To Reactive Power/Voltage Control In A Distribution Substation. IEEE Transactions on Power Systems Pwrs, 1997, 12(2): 681～688

**校对报告**

当前使用的样式是 [清华大学学位论文]

当前文档包含的题录共15条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常

**校对报告**

当前使用的样式是 [清华大学学位论文]

当前文档包含的题录共0条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常

**校对报告**

当前使用的样式是 [清华大学学位论文]

当前文档包含的题录共0条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常

**校对报告**

当前使用的样式是 [清华大学学位论文]

当前文档包含的题录共0条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常

**校对报告**

当前使用的样式是 [清华大学学位论文]

当前文档包含的题录共0条

有0条题录存在必填字段内容缺失的问题

所有题录的数据正常