基于深度强化学习的电梯智能调度系统

王丽冰　马仕君　徐素洁　孙琳琳　韩　康　胡　欣
北京邮电大学电子工程学院

摘要　针对电梯环境的动态性和复杂性，以及多个目标之间的互相约束性，提出一种基于深度强化学习的电梯调度方法。针对电梯状态空间的维度灾难带来的策略优化难题，采用近端策略优化算法来降低训练的复杂度。仿真结果表明，所提算法可以以极低的复杂度实现电梯的优化控制，解决电梯的高效率运行与用户等待时间公平性间的矛盾。

关键词　智慧城市　电梯调度　多目标优化　深度强化学习　近端策略优化算法

1　引言

随着智能建筑的发展和社会的进步，人们对电梯提出了越来越高的要求。对电梯要求已从简单的响应楼层、轿厢召唤，发展为能耗尽量少、候梯时间尽量短、乘梯时间尽量短等多个要求。如何对多台电梯进行有效的调度管理，提高电梯服务效率和质量是人们不得不面对的一个重要课题[1-4]。

电梯是在连续时间和空间上运行的动态系统，请求乘梯的用户到达是一个随机过程。传统的电梯调度方法，例如专家系统[5]、模糊控制[6-8]、神经网络[9]、遗传算法[10]等，由于分别具有依赖专家、不具有学习功能、学习不充分和不具有实时性的缺点，难以很好地应对复杂的电梯环境。强化学习通过在智能体与环境的交互过程中不断学习策略以达成回报最大化，从而实现特定目标，AlphaGO[11]的成功使其成为研究热点。近年来强化学习已广泛应用到机器人行为规划、游戏[12]、资源管理[13-15]、各种控制场合和调度问题[16]中。对于电梯环境的动态性和随机性，深度强化学习[17]通过将具有实时决策能力的强化学习与具有感知能力的深度学习相结合，可以实现对电梯的动态智能调度，提高电梯运行效率和服务质量。基于此，本文将电梯调度问题建模为序列决策过程，运用深度强化学习对电梯运行进行智能决策。近年来已有一些学者将深度Q网络（Deep Q Network，DQN）[18-20]应用到电梯调度问题中，与传统的方法相比，该方法能够应对未知的人流分布，但是复杂度比较高。针对此问题，本文采用近端策略优化方法（Proximal Policy Optimization，PPO）进行电梯调度，可以以极低的复杂度实现对电梯的动态调度。同时本文综合考虑用户等待时间、电梯能量消耗和用户服务公平性，将电梯调度视为一个多目标决策问题，从而使电梯综合效率达到最优。

2　系统模型

电梯调度系统是指控制中心对某一高层大楼内的多台电梯进行协同统一派梯调度。当大楼内某一楼层产生呼梯信号时，电梯调度系统会收集大楼内各台电梯的运行信息，根据某种派梯规则充分权衡用户的乘候梯时间、电梯能量消耗等各项指标，做出最佳派梯方案，选择最合适的电梯去响应该楼层的呼梯信号。整个电梯调度系统一般由单梯控制器、信号传输系统和电梯智能控制终端组成，如图1所示。信号传输系统将乘客的呼梯信号传入电梯智能控制终端。由电梯智能控制终端对电梯进行分派，并将结果传输给各单梯控制器；单梯控制器根据收到的派梯请求对电梯进行运行控制，从而实现对各电梯的协调控制。

本文模拟一个具有n部电梯的m层大厦，n部电梯由一个控制终端统一调度，每层提供一组呼梯按钮，用于候梯用户请求电梯的到达。每部电梯内部

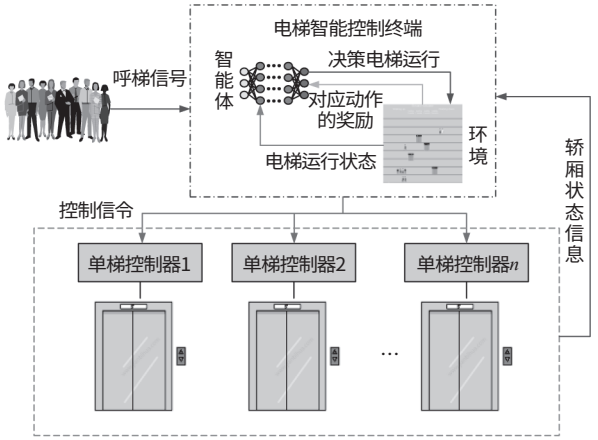

图1　电梯调度系统

提供一个控制面板，用于乘梯用户对目标楼层的选择，并显示电梯当前所处楼层和运行方向。电梯交通流作为电梯调度系统中的呼梯请求，是由电梯系统中用户数量、用户出现的周期以及用户的分布情况来描述的。电梯调度系统的交通流是一个随机过程，通常情况下，其可以看作一个泊松过程，根据泊松分布随机生成用户。

电梯在运行过程中，时刻需要观测系统状态，包括电梯本身状态和到达人流状态。当有新人流到达时，需要响应用户请求，即系统根据请求用户所在楼层分派电梯将用户从源楼层送至目标楼层，包括电梯抵达请求源楼层、电梯开关门、用户乘梯、电梯抵达目标楼层等过程。电梯系统运行逻辑流程如图2所示。

电梯作为一种智慧服务系统，其设计不仅应该考虑用户的服务质量，还应该考虑经济效益。从服务质量角度来说，人们总是希望候梯时间和乘梯时间越短越好；从节约能耗角度来说，电梯应避免空驶，减少起停次数。根据这些特点，电梯智能控制的目的就是得到一种派梯方案，使其能最好满足多个控制目标的要求，即最小化用户等待时间和电梯能量消耗。除此之外，电梯调度系统应该在减小用户平均等待时间的同时，尽可能地避免出现用户迟迟等不到电梯的恶劣情况。因此除了上述两个指标，用户等待时间公平性也应该作为优化目标。可以将电梯调度问题建立为一个多目标优化问题，如式(1)所示

$$\text{opt.min}(\lambda_1 \times T_t+\lambda_2 \times E_t+\lambda_3 \times G_t) \tag{1}$$

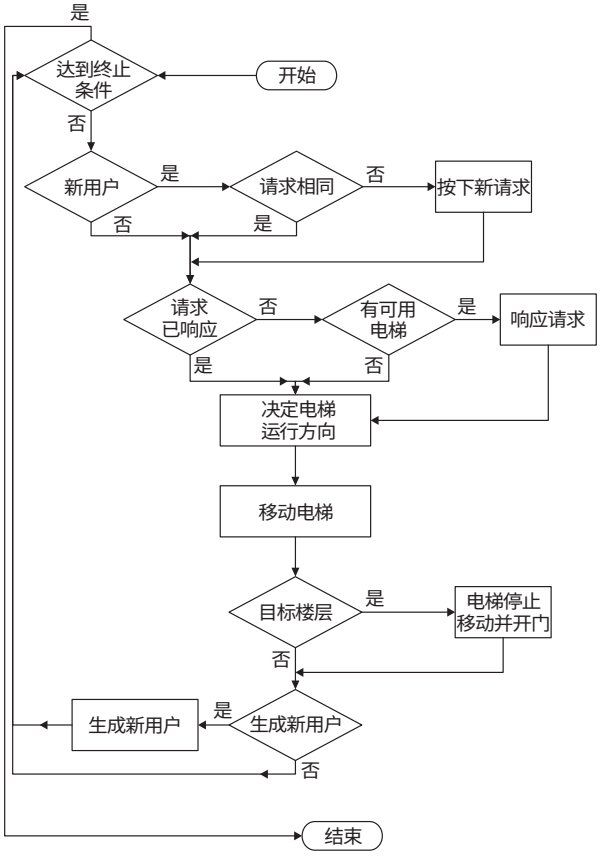

图2　电梯系统运行逻辑流程

其中，T_t为用户等待时间；E_t为电梯能量消耗，假设用户等待时间超过2min用户自动放弃等待；G_t为单位时间内放弃等待的用户个数，λ_1、λ_2和λ_3分别为以上3部分的权重。考虑到量纲不同，本文设置$\lambda_1=-1$，$\lambda_2=-5\times10^{-4}$，$\lambda_3=-300$。

（1）用户平均等待时间

在这里，用户等待时间包括候梯时间和乘梯时间，用户在时刻t的平均候梯时间通过式(2)计算

$$T_{1t}=\frac{1}{p_t}\sum_{i=1}^{p_t}t_{1t}(i) \tag{2}$$

其中，p_t表示当前时刻的候梯用户数量；$t_{1t}(i)$表示第i个用户的候梯时间。

用户在时刻t的平均乘梯时间通过式(3)计算

$$T_{2t}=\frac{1}{p_t^{'}}\sum_{j=1}^{p_t^{'}}t_{2t}(j) \tag{3}$$

其中，$p_t^{'}$表示当前时刻的乘梯用户数量；$t_{2t}(j)$表示第j个乘客的乘梯时间。

用户在时刻t的平均等待时间的计算公式为式(4)

$$T_t=T_{1t}+T_{2t} \tag{4}$$

（2）电梯能量消耗

电梯总能耗的组成部分包括运行和待机时的能耗，轿厢、竖井和机房的照明能耗，轿厢和机房的通风空调能耗[21]。其中最主要的部分是运行过程中的能量消耗，它取决于乘客负载、电梯速度以及系统某些部件所造成的能量损失；可以根据电机花费的时间计算运行时的能量消耗，即将电功率乘以运行时间，其计算公式如式(5)所示

$$E_{1t}=P_{\text{electricity}}\times t_{\text{run}}/3600 \tag{5}$$

其中，$P_{\text{electricity}}$为电功率，t_{run}为电机运行时间，时间单位转换为小时。

除了上述能量之外，还包括由于门的打开和关闭而产生的能量消耗E_{2t}，因此总能量消耗应由式(6)计算

$$E_t=E_{1t}+E_{2t} \tag{6}$$

（3）丢弃用户数量

通常情况下，用户在候梯时间超过2min之后放弃等待，为尽量避免丢弃用户，保证服务公平性，本文将丢弃用户数量目标考虑在内，其计算公式如式(7)所示

$$G_t=\sum_{t_{1t}(i)>2\min} P_i \tag{7}$$

其中，p_i表示第i个用户。

3 算法设计

3.1 电梯调度问题建模

电梯调度系统是连续时间的动态系统，电梯智能控制终端以时间触发机制观测系统状态并进行决策计算。由于电梯调度问题在控制的观点上是序列决策问题，电梯系统在当前时刻的状态仅与上一时刻的状态有关，而与之前的状态无关，因此可以将该问题建模为马尔科夫决策过程（Markov Decision Process，MDP），如图3所示。

系统环境状态作为电梯智能控制终端派梯决策的依据，应该包括电梯运行信息，如电梯运行方向、当前所在位置以及承载用户数量等信息，还包括用户请求状态，如大厦中各层楼的外呼请求和电梯内部乘梯用户的目标楼层请求等信息。因此将电梯调度系统定义为这两个部分状态的集合，表示为式(8)

$$s_t=[s_t^e,s_t^p] \tag{8}$$

其中，s_t^e表示电梯运行状态，s_t^p表示用户请求状态。表1显示了电梯系统具体状态设置。

电梯调度系统通过当前系统状态进行派梯决策，动作定义为式(9)

$$a_t=[a_{1,t}^1,a_{1,t}^2,a_{2,t}^1,a_{2,t}^2,\cdots,a_{n,t}^1,a_{n,t}^2] \tag{9}$$

其中，n是电梯数量，$a_{n,t}^1$，$a_{n,t}^2(i=1,2,\cdots,n)$分别表示第$i$个电梯分派的目标楼层和运行方向。

动作执行后，系统转移到下一个状态，同时给出一个奖励函数用来评价对应动作的好坏。根据式(1)，将奖励函数定义为式(10)

$$R_t=-(T_t+5\times10^{-4}E_t+300\times P_t)\times10^{-4} \tag{10}$$

3.2 基于深度强化学习的电梯调度算法设计

针对电梯环境的动态性、复杂性和随机性，本文利用深度强化学习方法对电梯进行智能控制，深度强化学习的学习过程是智能体与环境的交互过程，如图4所示。对于电梯调度问题，将大厦和电梯看作环境，将电梯智能控制终端看作智能体，智能体通过定时观测电梯运行和呼梯状态，决策派梯行为，并根据收到的奖励函数不断调整派梯策略使累计回报函数最大，从而找到最优派梯策略。

针对电梯状态空间的维度灾难带来的策略优

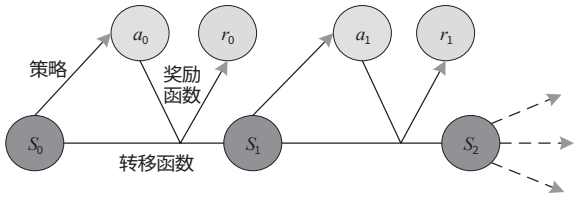

图3　马尔科夫决策过程

表1　系统状态设置

用户请求状态	电梯外有用户请求的楼层 电梯内用户的目标楼层
电梯运行状态	电梯当前所在楼层
	电梯当前速度
	电梯最大速度
	电梯当前方向
	电梯门当前打开的比例
	电梯当前承载的重量
	电梯最大能承载的重量
	电梯是否超载
	电梯门是否正在打开
	电梯门是否正在关闭

化难题，本项目采用近端策略优化算法（Proximal Policy Optimization，PPO）来实现电梯智能调度。近端策略优化算法是策略梯度方法的一种，非常适合应用在具有连续的动作和状态的空间问题中，而且该算法通过限制每步策略更新幅度的大小降低了策略梯度方法对参数的敏感度，能够获得更好的性能，同时具有较低的复杂度，因此本文采用该算法对电梯进行智能调度。近端策略优化算法本质上是一个演员—评论家（Actor-Critic）算法，其包含两个网络，其中决策网络即演员通过观测系统状态不断地做出决策与环境互动，同时价值网络即评论家针对环境对演员做的动作的反馈来评价动作的好坏，不断更新网络参数，使决策不断优化。除此之外，在该算法中，包含两个决策网络，分别为$\pi(\theta)$和$\pi(\theta_{old})$，其中$\pi(\theta_{old})$用来与环境互动，$\pi(\theta)$用来进行更新，在每个周期结束时将$\pi(\theta_{old})$更新为$\pi(\theta)$。其算法流程见表2。

其中iteration为周期，horizon为时间步，epoch为更新周期。T为每周期的训练步数，K为更新周期数。在基于近端策略优化算法的电梯调度系统中，智能体使用决策网络$\pi(\theta_{old})$与环境互动，采集样本存储在缓冲器中。与此同时，价值网络使用采集的样本对决策网络$\pi(\theta)$进行更新，并在周期结束时将$\pi(\theta_{old})$更新为$\pi(\theta)$，以在下一周期使用新的决策网络。如此迭代，不断对决策网络进行更新，直到找到最优策略。

4　仿真分析

为评估所提方法的性能，本文对电梯调度系统进行了仿真，模拟一个具有10层楼的大厦，该大厦具有4部电梯。系统参数和算法参数设置见表3。

针对上述参数构建电梯调度系统环境，采用深度强化学习方法中的近端策略优化算法对电梯进行智能调度，其训练过程如图5所示，其中横坐标是周期数，纵坐标是每个周期的累计奖励。由图5可以看出，奖励不断增加，并在1500个周期达到收敛，奖励值稳定在-5到-4，表示智能体在不断优化策略，在1500个周期时找到最优策略。加载该模型测试30个周期，得到用户等待时间、总电梯能量消耗和丢弃用户数量指标分别如图6～图8所示。

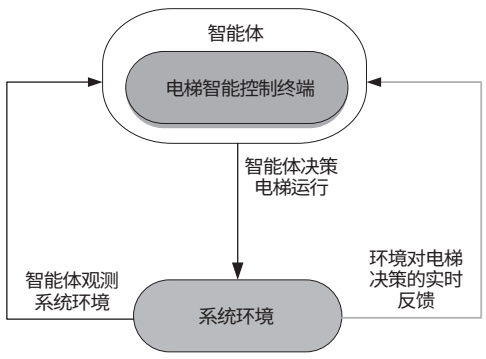

图4　深度强化学习

为了评估算法性能，实现了基于遗传算法（Genetic Algorithm，GA）的电梯调度。遗传算法作为一种全局优化算法，其在多目标优化方面具有良好的性能。由图8可以看到，在30个周期内，基于近端策略优化算法的电梯调度算法很少有丢弃用户的情况；而基于遗传算法的电梯调度算法几乎在每个周期内都有丢弃用户的现象，表示所提方法在保证用户等待时间的公平性方面有良好的性能。在图6中，基于近端策略优化算法的电梯调度算法用户累计等待时间明显低于基于遗传算法的电梯调度算法的用户累计等待时间，其控制在2500s左右。经统计，在20个周期内共出现用户数153个，因此用户平均等待时间为16.3s，这在实际生活中能满足大部分用户的需求。如图7所示，基于遗传算法的电梯调度算法电梯能量消耗在1250000J左右波动，而基于近端策略优化算法的电梯调度算法电梯能量消耗在1000000J左右，大大减小了电梯能量消耗。因此，本文所提算法使以上3个指标均达到了更好的效果，综合效率更优。

在复杂度方面，由于遗传算法是启发式方法的一种，当面对动态的电梯环境时，需要在每个时刻都进行迭代求解，造成很高的时间复杂度。而基于深度强化学习的方法，一旦模型训练得到最优策略，便可以对任意时刻状态给出一个即时的决策，大大降低了时间复杂度。为进一步评估近端策略优

表2　基于近端策略优化算法的电梯调度流程

1：根据系统参数生成电梯调度系统场景
2：初始化PPO算法网络，设置$\pi(\theta)=\pi(\theta_{old})$
3：对iteration=1,2,…循环:
4：对 horizon=1,2,…,T循环:
5：智能体观测系统状态s_t，包括电梯运行状态和用户请求状态
6：智能体使用$\pi(\theta_{old})$选择动作a_t，即进行派梯
7：环境反馈给智能体一个奖励函数r_t，即关于用户等待时间、电梯能量消耗和用户服务公平性的函数
8：将(s_t,a_t,r_t)存储到缓冲器
9：结束循环
10：对epoch=1,2,…, K循环:
11：从缓冲器中采样批量样本
12：通过梯度上升方法更新参数θ
13：结束循环
14：将$\pi(\theta_{old})$更新为$\pi(\theta)$

表3　仿真参数设置

参数		值
系统参数	开始时间	0
	时间间隔	0.5s
	楼层数	10层
	楼层高度	4m
	电梯数量	4个
	最大加速度	1.0m/s^2
	最大速度	2.0m/s
	人进电梯时间	2.0s
	开关门时间	2.0s
	自动门电源	350W
	待机功耗	100W
	最大容量	16人
	最大并行进出人口数	2人
	额定负荷	600
	净重	300kg
	滑轮半径	0.27m
	电机效率	0.8kW
	电机齿轮比	1.0
	齿轮效率	1.0
近端策略优化算法参数	学习率	10^{-4}
	折扣因子	0.99
	截断常数	0.2
	每周期的步数	3600步
	样本大小	64/k
	每周期更新次数	10/次
遗传算法参数	迭代次数	50/次
	种群数量	50/个
	染色体长度	4
	交叉概率	0.6
	变异概率	0.01

化算法在复杂度方面的性能，本文实现了基于深度Q网络的电梯调度仿真，其训练过程如图9所示。由图9可以看到，奖励值在1亿步才开始收敛，且最终可以达到−8。与基于深度Q网络的算法相比，近端策略优化算法收敛速度快、复杂度低，且最终达到的效果更好。因此，本文所提算法，不仅提高了系统性能，而且大大降低了复杂度。

5　结束语

采用深度强化学习方法对电梯进行智能调度，

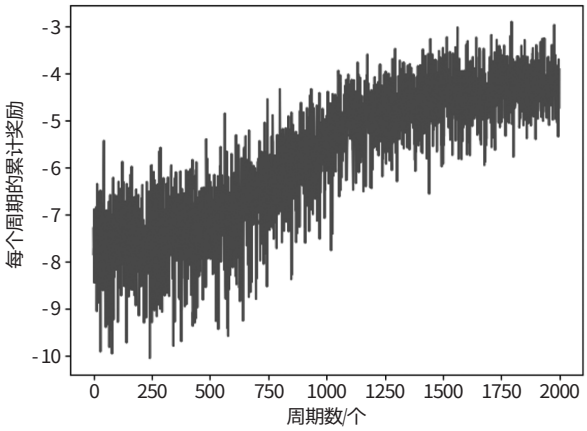

图5　PPO训练过程

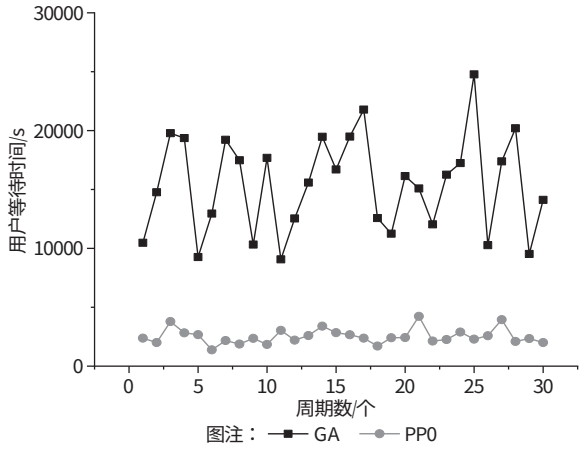

图6　用户累计等待时间

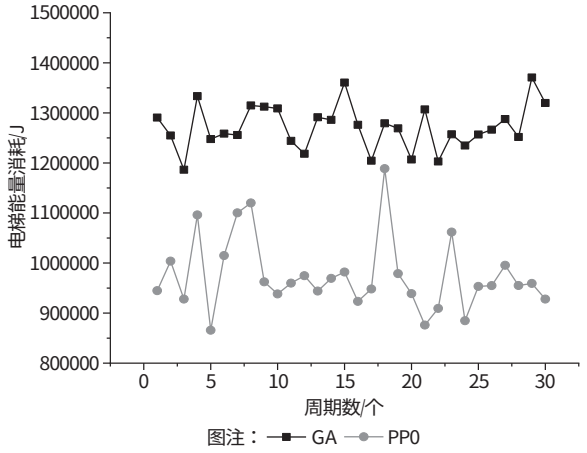

图7　总电梯能量消耗

通过智能体与电梯环境的互动不断优化电梯策略来实现用户等待时间最少、电梯能量消耗最小和用户服务公平性的最佳调度。同时，针对电梯状态空间

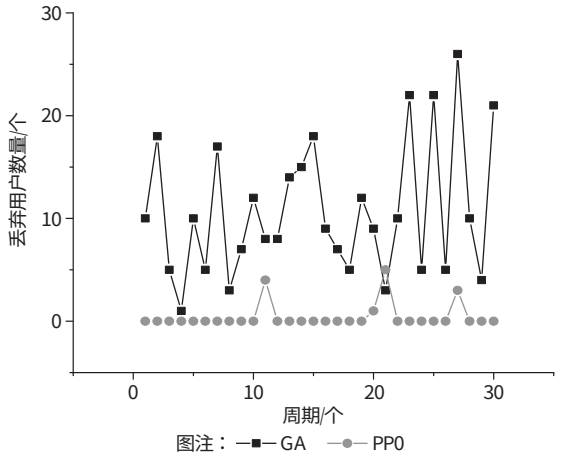

图8　丢弃用户数量

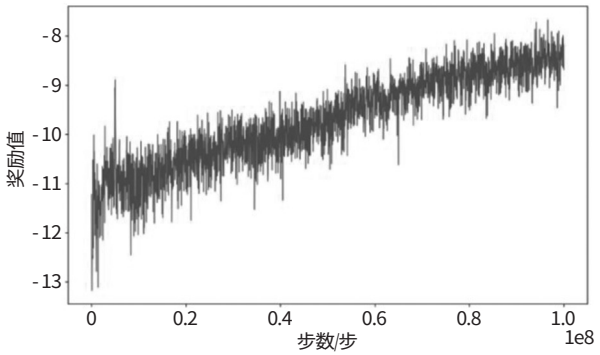

图9　DQN训练过程

的维度灾难带来的策略优化难题，本文采用近端策略优化算法，大大降低了训练复杂度。

参考文献

[1] Brand M, Nikovski D. Optimal parking in group elevator control[C]//Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on. Piscataway: IEEE Press, 2004.

[2] Crites R H, Barto A G. Elevator group control using multiple reinforcement learning agents[J]. Machine Learning, 1998, 33(2-3):235-262.

[3] Tervonen T, Hakonen H, Lahdelma R. Elevator planning with stochastic multicriteria acceptability analysis[J]. Omega, 2008, 36(3):352-362.

[4] Lee K, Kang K C, Koh E, et al. Domain-Oriented engineering of elevator control software[C]// Conference on Software Product Lines: Experience & Research Directions: Experience & Research Directions. Kluwer Academic Publishers, 2000.

[5] Tsuji S, Amano M, Hikita S. Application of the expert system to elevator group-supervisory control[C]// Conference on Artificial Intelligence Applications. Piscataway: IEEE Press, 1989.

[6] Kim C B, Seong K A. Design and implementation of a fuzzy elevator group control system[J]. Systems Man & Cybernetics Part A Systems & Humans IEEE Transactions, 1998, 28(3):277-287.

[7] Jamaludin J, Rahim N A, Hew W P. An elevator group control system with a self-tuning fuzzy logic group controller[J]. IEEE Transactions on Industrial Electronics, 2010, 57(12):4188-4198.

[8] 周海丹, 赵国军, 徐雷. 基于模糊逻辑的预约电梯群控算法[J]. 机电工程, 2010(9):37-41.

[9] 弓箭, 刘强, 刘剑. 人工智能在电梯群控系统中的应用[J]. 沈阳建筑工程学院学报（自然科学版）, 2002, 10(4): 306-308.

[10] 何万里, 李桂芝, 钱伟懿. 改进的遗传算法在电梯群控中应用[J]. 渤海大学学报(自然科学版), 2007, 28(1):46-50.

[11] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676):354-359.

[12] 陈兴国, 俞扬. 强化学习及其在电脑围棋中的应用[J]. 自动化学报, 2016, 42(5):685-695.

[13] Hu X, Liao X, Liu Z, et al. Multi-agent deep reinforcement learning-based flexible satellite payload for mobile termi-nals[J]. IEEE Transactions on Vehicular Technology, 2020(99): 1.

[14] Xin H, Shuaijun L, Rong C, et al. A deep reinforcement learn-ing-based framework for dynamic resource allocation in multibeam satellite systems[J]. IEEE Communications Letters, 2018, 22:1612-1615.

[15] Liao X, Hu X, Liu Z, et al. Distributed intelligence: A verification for multi-agent DRL based multibeam satellite resource allocation[J]. IEEE Communications Letters, 2020 (99):1.

[16] 李大字, 褚建华, 靳其兵. 基于强化学习算法的电梯群控系统仿真研究[C]// 第19届中国过程控制会议. 北京:化学工业出版社,2013.

[17] Li Y. Deep reinforcement learning: An overview[J]. 2017.

[18] Osband I, Blundell C, Pritzel A, et al. Deep exploration via bootstrapped DQN[J]. 2016.

[19] Zhang Q, Lin M, Yang L T, et al. Energy-efficient scheduling for real-time systems based on deep Q-Learning model[J]. IEEE Transactions on Sustainable Computing, 2017:1.

[20] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. 2017.

[21] Adak, M. Fatih, Nevcihan Duru, et al. Elevator simulator design and estimating energy consumption of an elevator system[J].Energy and Buildings, 2013 (65): 272-280.