

文章编号: 2095-4980(2023)01-0112-07

智能电网中基于增强学习的动态价格优化算法

曹 俊, 孙莹莹, 赵 航

(国网河南省电力公司 驻马店供电公司, 河南 驻马店 463000)

摘 要: 动态的电费价格是驱使消费者改变用电消费模式的有效手段, 为此, 提出基于增强学习的动态价格优化(RLODP)算法。RLODP 算法结合电力服务商的利润和消费者的用电成本, 对电网负载进行管理; 利用增强学习算法, 电力服务商自适应地决策零售价格, 将动态价格问题转化为离散有限马尔可夫决策过程(MDP), 再利用 Q-学习算法解决该决策过程。实验结果表明, 提出的 RLODP 算法减少了消费者的用电成本, 实现了电网市场中电力供应与需求之间的平衡。

关键词: 智能电网; 动态需求; 电价; 增强学习; 离散有限马尔可夫决策过程

中图分类号: TP393

文献标志码: A

doi: 10.11805/TKYDA2020178

Reinforcement Learning-based Optimizing Dynamic Pricing algorithm in smart grid

CAO Jun, SUN Yingying, ZHAO Hang

(Zhumadian Power Supply Company, State Grid Henan Electric Power Company, Zhumadian Henan 463000, China)

Abstract: Dynamic pricing is one of the most effective ways to encourage customers to change their consumption pattern. Therefore, Reinforcement Learning-based Optimizing Dynamic Pricing(RLODP) algorithm is proposed for energy management in a hierarchical electricity market by considering both service provider's profit and customers' costs. Using Reinforcement Learning, the SP can adaptively determine the retail electricity price. Dynamic pricing problem is formulated as a discrete finite Markov Decision Process(MDP), and Q-learning is adopted to solve this decision-making problem. Simulation results show that the RLODP algorithm can reduce energy costs for customers, balance the energy supply and the demands in the electricity market.

Keywords: smart grid; demand response; electricity price; Reinforcement Learning; discrete finite Markov Decision Process

为保证服务质量以及电网可靠性, 须有效地管理电网负荷, 电网负荷的增加可能毁坏电网设备。随着通信技术的发展, 先进读表基础建设(Advanced Metering Infrastructure, AMI)系统^[1]为提高用电效率、电网传输的可靠性以及能量效力提供了可能^[2-3]。

作为能源系统, AMI在电力供电站(供应商)与用电户(消费者)间实现双向通信, 其依据先进设备, 并结合网络, 对电网进行实时监测和管理。

双向通信模式便捷了供应商与消费者间的通信, 使它们能够交互一些必要的信息, 为实施需求响应(Demand Response, DR)策略提供了基础。通过收集消费者信息, 供应商对网络负荷进行管控, 进而调整电费价格。由于 DR 策略驱使了消费者对他们的用电进行管理, DR 策略已成为减少用电成本和提高电网可靠性的有效手段。DR 策略的实施可以使供电商与消费者间达到双赢: 消费者减少了用电成本, 供电电降低管理成本^[4]。

动态价格是 DR 策略中最有效的工具手段之一^[5], 能够激励消费者依据电网成本调整他们的用电, 使消费者错峰用电, 降低用电成本。然而, 每个消费者用电量的不确定性、消费者数量的波动等因素, 给管理电网负荷提出了挑战。人工智能(Artificial Intelligence, AI)技术的兴起, 为高效管理电网负荷赋予了新希望^[6]。如, 文献[7]将负荷服务企业(Load Serving Entity, LSE)与消费者间的电量交易看作离散-时间系统处理。LSE 通过增强学习(Reinforcement Learning, RL)设定用电的零售价, 进而最小化 LSE 和消费者的成本。

收稿日期: 2020-04-28; 修回日期: 2020-10-01

作为 RL 主流算法之一，Q-学习算法已在多个领域内广泛使用。Q-学习算法通过迭代更新动作价值函数，获取最优的 Q 值。为此，提出基于增强学习的动态价格优化(RLODP)算法。RLODP 算法将电价问题转化为离散有限马尔可夫决策过程(MDP)，利用 Q-学习算法求解，获取最优定价，进而实现在保护 LSE 的利润的同时，降低消费者的用电成本。仿真结果表明，提出的 RLODP 算法有效减少了消费者的用电成本。

1 RLODP 算法

1.1 系统模型

考虑如图 1 所示的系统模型，主要由 LSE 和 n 个消费者组成。令 θ_i 表示第 i 个用电消费者，且 $i=1,2,3,\dots,n$ 。在时刻 t ，消费者 θ_i 向 LSE 请求所需的能量 $E_i(t)$ 。收到请求后，LSE 就依据定价策略，设定电费价格，进而通过电费价格驱使消费者调整他们的用电量。

1.2 消耗者用电模型

依据用户用电的特点^[8]，将用户划分为必须满足用电户 (Critically Met User, CMU) 和弹性用电户 (Curtable User, CU)。LSE 必须满足 CMUs 所请求的用电量，即：

$$e_i^{\text{CMU}}(t) = E_i^{\text{CMU}}(t) \quad (1)$$

式中： $E_i^{\text{CMU}}(t)$ 表示用户 θ_i 请求的用电量； $e_i^{\text{CMU}}(t)$ 表示用户 θ_i 实际上消耗的用电量。

由于 CU 会因为电费价格的增加而调整用电量，不失一般性， $e_k^{\text{CU}}(t) \leq E_k^{\text{CU}}(t)$ 。因此，CU 用户 θ_k 在时隙 t 所消耗的电量 $e_k^{\text{CU}}(t)$ ：

$$e_k^{\text{CU}}(t) = E_k^{\text{CU}}(t) \left(1 + \xi_t \frac{r_k(t) - p(t)}{p(t)} \right) \quad (2)$$

式中： $E_k^{\text{CU}}(t)$ 表示用户 θ_k 所请求的用电量； $e_k^{\text{CU}}(t)$ 表示 LSE 所给予用户 θ_k 的用电量； ξ_t 表示在时隙 t 的弹性系数^[9]，且 $\xi_t < 0$ ； $r_k(t)$ 表示在时隙 t 的用户 θ_k 的零售用电价格； $p(t)$ 表示在时隙 t 的用户 θ_k 的批量用电价格，且 $r_k(t) \geq p(t)$ 。

RLODP 算法旨在控制 CU 用户的用电量，减少用户 θ_k 的成本。令 $C_k(t)$ 表示其在时隙 t 所减少的用电成本：

$$C_k(t) = \frac{\alpha_k}{2} (E_k^{\text{CU}}(t) - e_k^{\text{CU}}(t))^2 + \frac{\beta_k}{2} (E_k^{\text{CU}}(t) - e_k^{\text{CU}}(t)) \quad (3)$$

式中： α_k, β_k 分别表示用户 θ_k 的参数，且均大于零。此外，用变量 $E_{\max}$ 、 $E_{\min}$ 对 $(E_k^{\text{CU}}(t) - e_k^{\text{CU}}(t))$ 减少的用电量进行限定，即 $E_{\min} < (E_k^{\text{CU}}(t) - e_k^{\text{CU}}(t)) < E_{\max}$ 。

最后，任意用户 θ_i 在时隙 t 所花费的用电成本为：

$$G_i(t) = r_i(t) [f_i^{\text{CMU}} e_i^{\text{CMU}}(t) + f_i^{\text{CU}} e_i^{\text{CU}}(t)] + f_i^{\text{CU}} C_i(t) \quad (4)$$

式中 $f_i^{\text{CMU}}, f_i^{\text{CU}}$ 为布尔变量。若用户 θ_i 为 CMU，则 $f_i^{\text{CMU}}=1, f_i^{\text{CU}}=0$ ；否则， $f_i^{\text{CMU}}=0, f_i^{\text{CU}}=1$ 。因此， $f_i^{\text{CMU}} + f_i^{\text{CU}}=1$ 。

提出 RLODP 算法的目的在于降低在观察时隙内的用电成本，因此，可建立目标函数：

$$\min \sum_{t=1}^T \sum_{i=1}^n G_i(t) \quad (5)$$

式中 T 为时隙总数。

1.3 LSE 供电模型

LSE 从电网企业 (Grid Operator, GO) 以批量价格 $p(t)$ 购买电量，然后以零售价向消费者出售，进而实现赢利^[10]。对于任意 LSE，利润最大化是其目标：

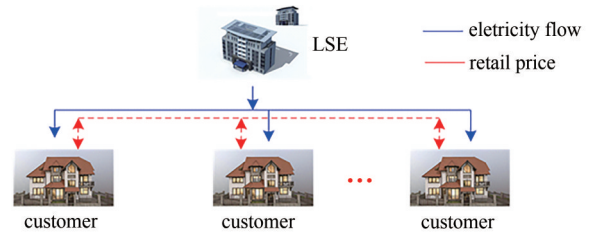


Fig.1 System model
图 1 系统模型

$$\max \sum_{i=1}^N \sum_{t=1}^T \{ [r_i(t) - p(t)] [f_i^{\text{CMU}} e_i^{\text{CMU}}(t) + f_i^{\text{CU}} e_i^{\text{CU}}(t)] \} \quad (6)$$

通常, $r_i(t)$ 大于 $p(t)$, 但对它们的差值是有约束的。

1.4 目标函数

设计 RLODP 模型的目的是实现消费者与 LSE 的双赢, 即减少消费者的用电成本和最大化 LSE 的利润。为此, 可建立目标函数:

$$\max \sum_{i=1}^N \sum_{t=1}^T [F_i(t)] \quad (7)$$

式中: $F_i(t) = \rho[r_i(t) - p(t)]e_i(t) - (1 - \rho)[r_i(t)e_i(t) + C_i(t)]$, 其中 $e_i(t) = f_i^{\text{CMU}} e_i^{\text{CMU}}(t) + f_i^{\text{CU}} e_i^{\text{CU}}(t)$, $\rho \in [0, 1]$ 表示消费者的用电成本与 LSE 的利润间的权重因子, 其值由 LSE 决定。

2 基于 RL 算法的双赢模型

引用基于 RL 算法的双赢模型, 如图 2 所示。LSE 作为代理, 消费者为环境, 零售价格表述动作, 即 LSE 在每个时隙向消费者反馈零售用电价格; 消费者的用电信息, 包括用电需求和实际的用电量代表状态。为此, 将动态的零售价格优化问题转化为离散的有限马尔可夫决策过程(MDP); 再结合 Q-学习算法, 提出有效的动态价格 DR 算法。

2.1 MDP 参数的初始化

MDP 的关键参数包括: 离散时间 t ; 动作 $A(r_i(t))$; 状态 $S(E_i(t), e_i(t))$; 奖励 $R(\zeta(e_i(t)|E_i(t), r_i(t)))$ 。对于 MDP 的每一轮, 可计算总的奖励:

$$R = \sum_{i=1}^n [\zeta(e_i(1)|E_i(1), r_i(1)) + \zeta(e_i(2)|E_i(2), r_i(2)) + \dots + \zeta(e_i(T)|E_i(T), r_i(T))] \quad (8)$$

式中: $\zeta(e_i(t)|E_i(t), r_i(t)) = F_i(t)$ 。用 R_t 表示 t 时隙后的奖励, 包括 t 时隙, 其定义如式(9)所示:

$$R_t = \sum_{i=1}^n [\zeta(e_i(t)|E_i(t), r_i(t)) + \zeta(e_i(t+1)|E_i(t+1), r_i(t+1)) + \dots + \zeta(e_i(T)|E_i(T), r_i(T))] \quad (9)$$

考虑未来环境的不确定性, 对未来奖励采用折扣因子, 如式(10)所示:

$$R_t = \sum_{i=1}^n [\zeta(e_i(t)|E_i(t), r_i(t)) + \gamma \zeta(e_i(t+1)|E_i(t+1), r_i(t+1)) + \dots + \gamma^{T-t} \zeta(e_i(T)|E_i(T), r_i(T))] \quad (10)$$

式中 γ 为折扣系数, 且 $\gamma \in [0, 1]$ 。根据状态与动作的映射关系定义定价政策: $\mathfrak{S}: r_i(t) = \mathfrak{S}(E_i(t))$ 。提出 RLODP 模型的目的就是寻找最优的定价政策 $\mathfrak{S}$, 即选择最优的动作(零售价格), 使消费者与 LSE 间达到双赢。

2.2 基于 Q-学习的定价政策

2.2.1 Q-学习的改进

作为免模型的 RL 技术, Q-学习常被利用获取最优的政策。设计 Q-值 $Q(\zeta(e_i(t)|E_i(t), r_i(t)))$ 是 Q-学习的基本步骤。即在每个时隙 t , 为每个状态-动作构建 Q 值。然后, 再通过迭代更新, 进而获取最优的动作。引用式(11)更新 Q 值:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (11)$$

式中: α 为学习率, 决定了此时 Q 值接收目标 Q 值的近似程度, 且 $\alpha \in [0, 1]$ 。若 $\alpha = 0$, 表示学习停止。在确认环境中, 可以使 α 取值为 1; γ 为折扣因子, 当 $\gamma = 0$, 表示学习体(Agent)只考虑当前最优的回报; 当 γ 逐步接收

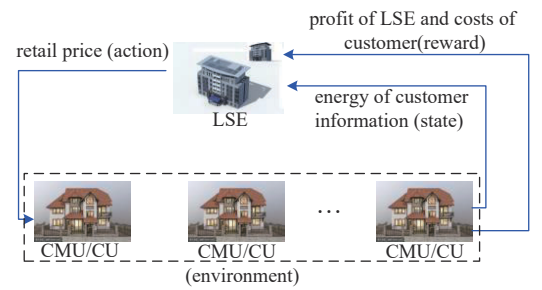


Fig.2 Win-win model based on RL
图2 基于RL的双赢模型

1 时, 学习体也逐步考虑高的回报值^[11]; S_t, A_t 分别表示时隙 t 的状态、动作, 其中 $S_t = S(E_i(t), e_i(t))$, $A_t = A(r_i(t))$; R_{t+1} 表示时隙 $t+1$ 的奖励, 即 $R_{t+1} = R(\zeta(e_i(t+1)|E_i(t+1), r_i(t+1)))$ 。为了简化表述, 用 S_t, A_t 和 R_{t+1} 分别表示时隙 t 的状态、动作和奖励。

为防止陷入局部最优解, 对 Q-学习算法进行改进。在训练初期阶段引入 ε -贪婪算法, 使学习体以一定的概率对环境进行探索。通过经验的逐步积累, 获取稳定收敛的 Q 值。

针对未知环境, Agent 通过不断试错学习, 积累经验。第一步: 依据 ε -贪婪算法选择动作并执行; 状态转移, 并从环境获取奖赏; 第二步: 将奖赏值传递给 Q 值, 并进行更新; 第三步: 将多次训练后的 Q 值存储于 Q 表格中, 再从此表格中获取最优的动作, $A^* = \arg \max_{A \in A} Q(S, A)$, 其中 A 为动作集。重复上述过程, 直到 Q 值收敛, 进而获取一个最优的策略。当 Agent 学习完成后, 就将 ε -贪婪算法中的 ε 值设置为零。

图 3 给出了 Agent 基于改进 Q-学习算法选择动作的过程。由于 Agent 在学习过程中, 可能会遭遇负奖赏, 使 Agent 不会复位到最初位置, 主要原因是最初 Agent 学习环境的经验不足, 为一个训练周期(episode)积累的知识不充裕, 减少了探索步数, 延缓了学习进度。因此通过改进算法, 弥补初期经验缺乏的不足, 提升 Q 值收敛速度, 缩短训练周期。

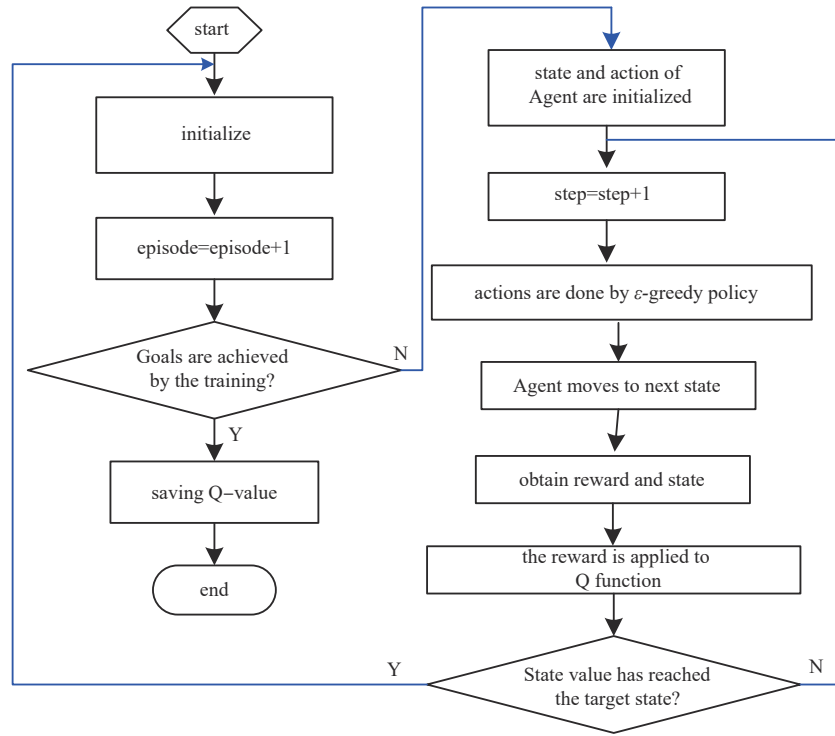


Fig.3 Action is controlled by improved Q-learning of Agent

图3 学习体利用改进后的 Q-学习算法控制动作流程图

2.2.2 基于改进后 Q-学习算法的定价

令 $Q^*(\zeta(e_i(t)|E_i(t), r_i(t)))$ 表示最优的 Q 值, 表示在状态 $E_i(t)$ 时, 采用动作 $r_i(t)$ 时获取最优的利润。通过满足贝尔曼方程(Bellman Equation, BE)^[12], 计算 $Q^*(\zeta(e_i(t)|E_i(t), r_i(t)))$ 值:

$$Q^*(\zeta(e_i(t)|E_i(t), r_i(t))) = \zeta[e_i(t)|E_i(t), r_i(t)] + \gamma \max Q \{ \zeta[e_i(t+1)|E_i(t+1), r_i(t+1)] \} \quad (12)$$

图 4 为 RLODP 算法的流程:

第一步: 初始化, 获取消费者对用电的需求 $E_1(t), E_2(t), \dots, E_n(t)$ 、参数 $\alpha_i, \beta_i, E_{\max}, E_{\min}$ 、批量价格 $p(1), p(2), \dots, p(T)$ 、弹性系数 ζ_i 、权重因子 ρ 。

第二步: 将 Q 值进行初始化 $Q(\zeta(e_i(t)|E_i(t), r_i(t))) \leftarrow 0$; 赋时隙值 $t \leftarrow 1$; 赋迭代次数 $k \leftarrow 1$;

第三步: 开始每次迭代, 先观察在时隙 t 所需的用电 $E_i(t)$, 然后通过 ε -greedy 机制^[13]计算零售价格 $r_i(t)$, 再计算奖励, 并估算 Q 值。

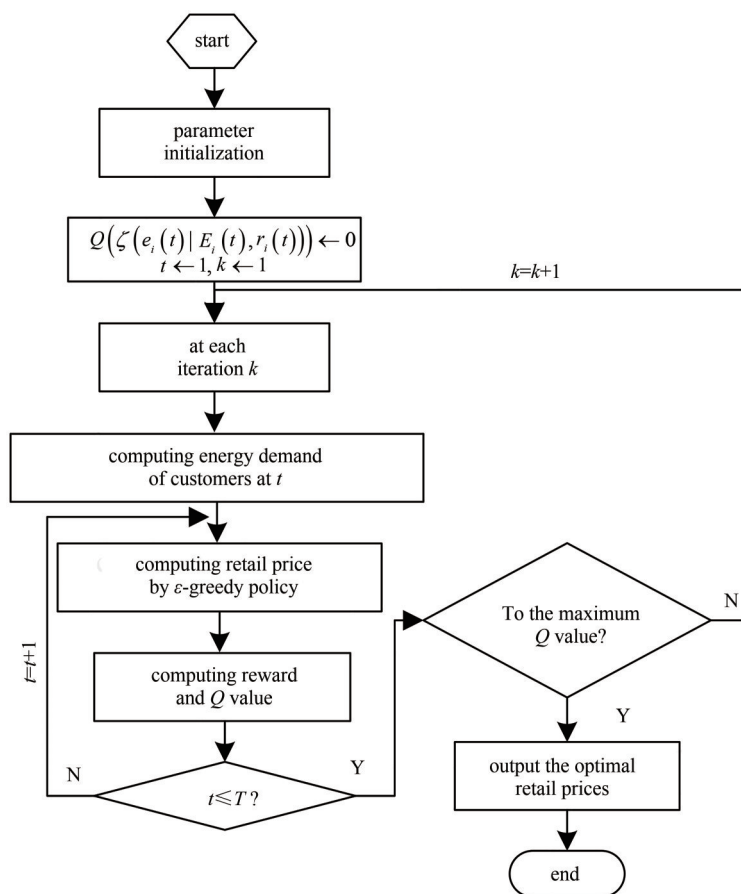


Fig.4 The flow of RLODP algorithm

图4 RLODP算法的流程

第四步：判断是否满足 $t \geq T$ ，如果满足，继续判断是否达到最优 Q 值，如果没有达到最优 Q 值，继续迭代，否则，得到最优的定价政策；若不满足 $t \geq T$ ，则增加时隙值，即 $t = t + 1$ ，继续迭代。

3 性能分析

3.1 实验参数

实验参数设置：3 个用户 ($n = 3$)；观察时隙 $T = 24$ ，即一天的时间，每个时隙的时长为 1 h。3 个用户的参数设置如下： $\alpha_1 = 0.8$ ， $\alpha_2 = 0.5$ ， $\alpha_3 = 0.3$ ； $\beta_1 = \beta_2 = \beta_3 = 0.8$ 。弹性系数 ξ_i 在 $-0.3 \sim -0.7$ 变化，其对应了一天的时隙，如表 1 所示。

表 1 不同时间段的弹性系数

Table 1 Elasticity coefficients at different time periods

	off-peak (morning: 1am—12am)	mid-peak (13pm—16pm, 22pm—24pm)	on-peak (17pm—21pm)
ξ_i	-0.3	-0.5	-0.7

3.2 用电量及零售价格性能

图 5 为 RLODP 算法 3 个用户零售价格和批量价格的设置情况。图 5 给出 24 个时隙内的用电情况、批量价格以及零售价格。

从图 5 可知，零售价格的趋势与批量价格类似，反映了 LSE 采购电能的成本。从 $t = 6$ 增至 $t = 12$ 时，消费者的零售价格让消费者获取更多的利益。但在 $t = 13$ 时，消费者的零售价格突然下降。原因在于： ξ_i 从 -0.3 变化至 -0.5 ，到达了近高峰时段。因此，零售价格的连续增加，加大了电能下降的量。

此外，观察图 5 不难发现，相比于用户 1 和用户 2，用户 3 的平均零售价得到提升。原因在于：用户 3 具有更

低的 α_3 值，减少了对能量所需的量，增加了平均零售价格。

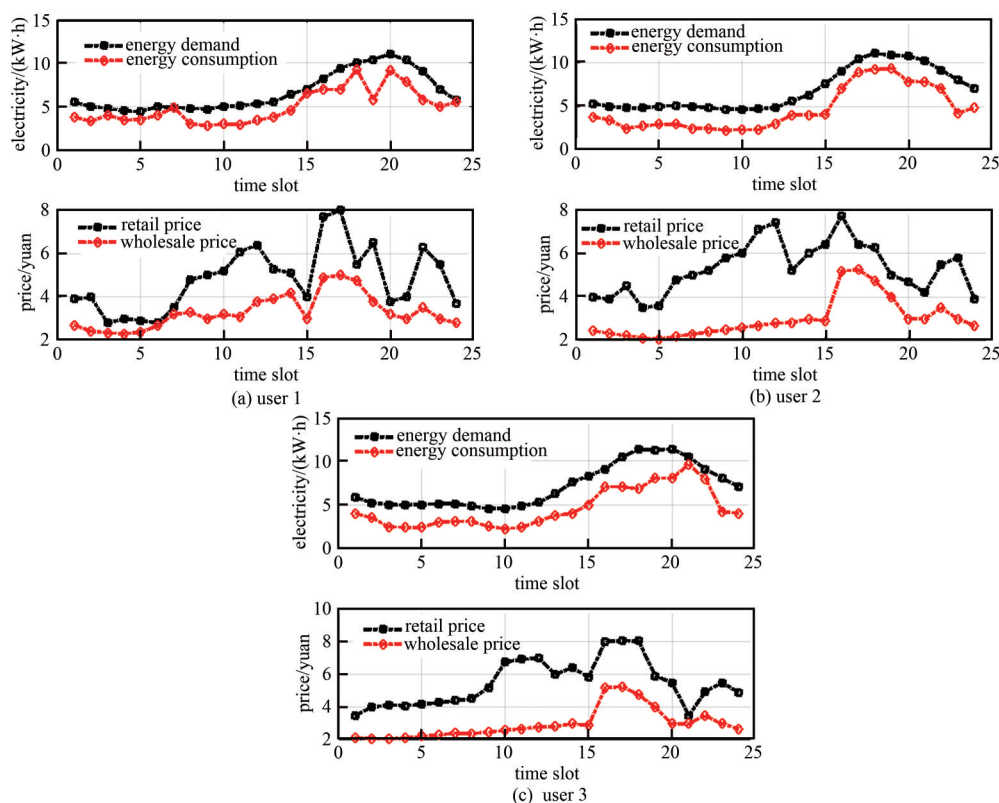


Fig.5 Energy consumption of three users at each time slot

图5 三个用户的用电情况

图6为3个用户在一天之内所消耗的总电量和其所请求的电量。从图6可知，用户1、用户2和用户3的请求电量与实用电量的差值分别为39.95 kW·h、52.79 kW·h和60.40 kW·h，即利用RLODP算法驱使用户减少了用电量。相比于用户2和用户3，用户1请求的电量与实用电量的差值更小。原因在于：用户的 α_1 值最大。 α_1 值越大，意味着用户所希望减少的用电量欲望越小。通过控制用户的用电量，平衡了电量供应与电量需求间的平衡，有效地管控了电网负荷，增加了电网系统的可靠性。

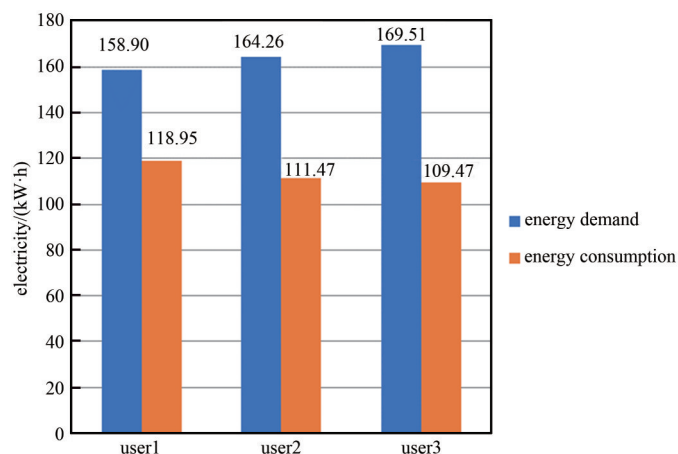


Fig.6 Energy demand and consumption of three users

图6 三个用户请求的电量和实用的电量

4 结论

针对电网价格优化问题，提出基于增强学习的动态价格优化RLODP算法。RLODP算法将定价问题进行MDP处理。依据消费者的需求以及LSE的利润，建立目标函数，并利用Q-学习算法获取最优的电费价格。实验数据

表明, 提出的 RLODP 算法能够驱使消费者调整用电, 实现对电网负荷的管控。

参考文献:

- [1] 张乐平,李向锋. 欧盟AMI经验对下一代IR46智能电表规划的启示[J]. 电测与仪表, 2019,56(22):146-152. (ZHANG Leping, LI Xiangfeng. EU-27 AMI rollout and inspiration of the IR46 smart meter planning[J]. Electrical Measurement & Instrumentation, 2019,56(22):146-152.)
- [2] MOHASSEL R R, FUNG A, MOHAMMADI F, et al. A survey on advanced metering infrastructure[J]. International Journal of Electrical Power & Energy Systems, 2014(63):473-484.
- [3] 田杰,程永生,肖何,等. 基于射频能量采集的Underlay CRN的能效优化[J]. 太赫兹科学与电子信息学报, 2020,18(3):397-404. (TIAN Jie, CHENG Yongsheng, XIAO He, et al. Energy efficiency optimization for Underlay Cognitive Radio Networks with RF energy harvesting[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(3):397-404.)
- [4] VARDAKAS J S, ZORBA N, VERIKOUKIS C V. A survey on demand response programs in smart grids: pricing methods and optimization algorithms[J]. IEEE Communications Surveys Tutorials, 2015,17(1):152-178.
- [5] SAMADI P, MOHSENIAN-RAD H, SCHÖBER R, et al. Advanced demand side management for the future smart grid using mechanism design[J]. IEEE Transactions on Smart Grid, 2012,3(3):1170-1180.
- [6] AGARWAL D, CHEN B C, ELANGO P, et al. Online models for content optimization[C]// Proceedings of the 21st International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: [s.n.], 2009:17-24.
- [7] GHASEMKHANI A, YANG L. Reinforcement learning based pricing for demand response[C]// 2018 IEEE International Conference on Communications Workshops (ICC Workshops). Kansas City, MO, USA: IEEE, 2018:1-6.
- [8] JIN M, FENG W, MARNAY C, et al. Microgrid to enable optimal distributed energy retail and end-user demand response[J]. Applied Energy, 2018(210):1312-1335.
- [9] LU Renzhi, HONG Seung Ho, ZHANG Xiongfeng. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach[J]. Applied Energy, 2018(220):220-230.
- [10] REZA B, NASSER M, BABAK B. Optimizing dynamic pricing demand response algorithm using reinforcement learning in smart grid[C]// 2020 25th International Computer Conference. Tehran, Iran: IEEE, 2020:1-5.
- [11] 徐娟. 基于强化学习的动作控制与决策研究[D]. 西安: 西安石油大学, 2020. (XU Juan. Research on action control and decision based on reinforcement learning[D]. Xi'an, China: Xi'an Shiyou University, 2020.)
- [12] 曹玉松. 哈密尔顿-雅可比-贝尔曼方程下的最优再保险策略[J]. 河南理工大学学报(自然科学版), 2016,35(2):285-288. (CAO Yusong. Optimal reinsurance strategy under Hamilton-Jacobi-Bellman equation[J]. Journal of Hehan Polytechnic University (Natural Science), 2016,35(2):285-288.)
- [13] 胡晓辉. 一种基于动态参数调整的强化学习动作选择机制[J]. 计算机工程与应用, 2008,44(28):29-31. (HU Xiaohui. Action choice mechanism of reinforcement learning based on adjusted dynamic parameters[J]. Computer Engineering and Applications, 2008,44(28):29-31.)

作者简介:

曹 俊(1981-), 男, 学士, 高级工程师, 主要研究方向为电气工程及其自动化 .email:huxyu_82@sohu.com.

赵 航(1987-), 男, 学士, 工程师, 主要研究方向为电气工程.

孙莹莹(1975-), 女, 硕士, 高级工程师, 主要研究方向为电气工程及其自动化.