MethodsX 12 (2024) 102735

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Application of deep reinforcement learning in electricity demand response market: Demand response decision-making of load aggregator ☆

Guangda Xu [a], Shihang Song [b,*], Yu Li [a], Yi Lu [a], Yuan Zhao [a], Li Zhang [b], Fukun Wang [b], Zhiyu Song [b]

[a] *State Grid Jibei Electric Power Research Institute (North China Electric Power Research Institute Co., Ltd), Beijing 100045, China*
[b] *Key Laboratory of Power System Intelligent Dispatch and Control of Ministry of Education (Shandong University), Jinan 250061, China*

ARTICLE INFO

Method name:
Demand response decision-making with dynamic classification of consumers

Keywords:
Demand response
Deep reinforcement learning
Load aggregator
Consumer classification
Electricity market

ABSTRACT

With the large-scale integration of renewable energy generation, developing the potential of demand response is becoming more and more significant. However, the number of consumers is huge, the electricity consumption behaviors are various and the information of consumers is incomplete, thus there are great difficulties in modeling and processing demand response by the conventional mathematical methods. Therefore, how to accurately predict the response behaviors of consumers and select the appropriate consumers for demand response is worthy of in-depth discussion. In this paper, the decision-making method of the load aggregators who participant in demand response market on behalf of the small and medium-sized consumers based on Deep Q-network algorithm is proposed, supporting the aggregators to properly guide the potential demand response consumers. At the same time, a dynamic consumer classification method based on self-organizing maps algorithm is also proposed, supporting the aggregators to accurately predict the response behaviors of the consumers. Simulation results show that the proposed method can effectively realize the classification of the consumers and result in a more beneficial demand response.

- The proposed method does not require complete information, such as demand response behavior parameters of the consumers.
- Self-organizing maps can realize a dynamic classification of demand response consumers as the case may be.
- DQN algorithm can effectively realize the demand response decision-making of aggregators under the condition of incomplete information.

☆ **Related research article:** Xu, G., Liu, L., Lu, Y., Zhao, Y., Zhang, L., & Song, S, Perception and decision-making for demand response based on dynamic classification of consumers. International Journal of Electrical Power & Energy Systems,148(2023), 108,954. https://doi.org/10.1016/j.ijepes.2023.108954

DOI of original article: 10.1016/j.ijepes.2023.108954

* Corresponding author.

E-mail address: songshhang@mail.sdu.edu.cn (S. Song).

https://doi.org/10.1016/j.mex.2024.102735
Received 2 October 2023; Accepted 25 April 2024
Available online 8 May 2024

Specifications table

Subject area:	Energy
More specific subject area:	Demand response
Name of your method:	Demand response decision-making with dynamic classification of consumers
Name and reference of original method:	N/A
Resource availability:	N/A

Method details

Introduction

Under the background of large-scale integration of renewable energy generation, the power balance of power system will face greater challenges. With the deepening of China's electricity market reform, the participation of demand-side resources in short-term electricity trading will become an inevitable trend to ensure the safe and reliable operation of power system and promote the mature development of the electricity market. Demand response can fully mobilize the enthusiasm of the flexible load on the demand side through price means, making flexible load quickly respond to the imbalance between supply and demand of power system, and greatly improve the reliability and economy of grid [1–2].

Previously, demand response optimization and decision-making problem are usually modeled as mathematical programming models, and can be classified as linear programming, mixed integer linear programming, nonlinear programming and other models according to the different forms of problems. Literature [3] uses linear programming to model household demand response, and seeks the optimal energy consumption of various electrical appliances in different time periods to achieve the optimization objective of minimizing household electricity cost and waiting time of household appliances. Literature [4] uses mixed integer linear programming to model the demand response of residential communities containing renewable energy power generation, and the optimization objective is the minimum energy consumption cost of residential communities. In literature [5], a mixed non-integer linear programming is used to model and schedule different electrical equipment in a family residence, with the optimization objective of minimizing energy consumption cost and maximizing electricity consumption comfort. Literature [6] uses nonlinear programming to realize the control strategies of different electrical equipment under different demand response forms.

However, because of the characteristics of flexible load on the demand side, such as the huge number, the scattered distribution in space and small monomer capacity, if traditional method based on mathematical programming mode is still used to control numerous consumers separately, the computational efficiency will be low and the scheduling cost will be high. And in real life, the electricity consumption behaviors of consumers are diverse, it is difficult to directly obtain the consumer's response behavior model parameters and carry out analysis based on the model directly. Therefore, it is necessary to aggregate the numerous dispersed consumers and explore a method that does not rely on the traditional mathematical model to aggregate and guide the demand response behavior of consumer.

As an important participant in electricity market, load aggregator (electricity selling company) has ability to integrate the demand response resources of small and medium-sized consumers. The framework of the two-layer transaction is shown in Fig. 1. On the one hand, the aggregator provides opportunities for small and medium-sized consumers to participate in demand response market and pursue benefits. On the other hand, the aggregators are professional, they can guide consumers to consume electricity more reasonably and provide demand response resources for the market. The existing literatures show that machine learning methods have certain advantages in studying the behavior of market participants. Literatures [7,8] use high-dimensional data perception ability of deep learning to predict the electricity consumption behavior of market participants, and literatures [9–13] use reinforcement learning to determine the optimal trading strategy of market participants. Deep reinforcement learning combines the decision-making advantage

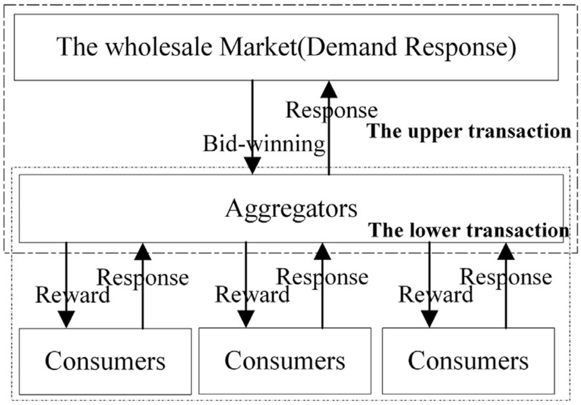


Fig. 1. Framework of the two- layer transaction mediated by the aggregator.

of reinforcement learning with the perceptual advantage of deep learning, enabling it to have the decision-making ability for high-dimensional data [14,15], and thus has better performance in real-time market transaction decision-making. For example, in literature [16], the pricing decision-making of wholesale purchase price and retail price in the electricity market is transformed into a Markov problem, and DDPG algorithm is used to solve the optimal strategy, which effectively increases the profits of the power producers. In literature [17], deep reinforcement learning is applied to determine the demand response of industrial consumers, and the actor-critic algorithm is used to determine the optimal energy management strategy. Literature [18] focuses on interruptible loads in electricity market and constructs an automatic demand response system based on DDQN algorithm. Literature [19] focuses on the optimization of power purchase decision-making, using deep neural network to predict price and demand, and using reinforcement learning method to solve the optimal decision-making.

In the demand-side resource transaction of real-time market, the uncertainty of market environment is difficult to describe by deterministic model. While deep reinforcement learning algorithm has advantages of model-free, perception of high-dimensional data and decision-making ability, which enables it to solve the optimal strategy under the condition of incomplete environmental information. Therefore, this paper proposes dynamic classification of numerous dispersed consumers through self-organizing maps (SOMs) algorithm, uses Deep Q-network (DQN) algorithm to select demand response consumers and determine the reward price sent to them. Finally, the dynamic classification of demand response consumers is realized and the aggregator can handle the optimal decision-making.

The remainder of this paper is organized as follows: Section 2 and 3 introduces the basic behavior of consumers and load aggregator. Section 2 introduces the behavior and data of the consumers participating in demand response. Section 3 introduces the decision-making behavior of the aggregator. The methods of decision-making of aggregator are proposed in section 4 and 5. Section 4 proposes the method of dynamic classification of consumer response behavior based on SOM. Section 5 proposes the decision-making method of the aggregator based on deep reinforcement learning. Simulation result and discussion of the proposed method are illustrated in section 6. Section 7 and 8 introduce the limitations and conclusion of this paper.

Behavior and data of the consumers participating in demand response

Modeling the consumer behavior in demand response

In demand response market, numerous dispersed consumers participate in market respected by the aggregators. The aggregator sends reward price $p_{xt,i}$ to the consumer *i*, and then the consumer adjusts his electricity consumption behavior according to the attainable reward, provides demand response volume $res_{t,i}$. However, due to the differences in electricity demand, preferences and sensitivity to electricity price among different classes of consumers [20], their actual response behaviors are various.

In order to effectively describe the relationship between the demand response volume $res_{t,i}$ of consumer *i* and the reward price $p_{xt,i}$ sent by aggregator, this paper defines a response behavior function of the consumers considering the different response willingness of the consumers, based on the quadratic function model of consumer response behavior proposed in previous research [21]. The definition is shown in Eq. (1).

$$\begin{cases} res_{t,i} = \alpha_{t,i}(p_{xt,i} - th_{t,i})^n \\ \alpha_{t,i} = \frac{q_{t,i}}{m} \end{cases} \tag{1}$$

where $res_{t,i}$ represents the response volume of the *i* th class consumer in period *t*. $p_{xt,i}$ represents the reward price sent by the aggregator to the *i* th class consumer in period *t*. $th_{t,i}$ represents the expected reward price of the *i* th class consumer in period *t*. If $p_{xt,i}$ is greater than $th_{t,i}$, the consumer will provide demand response. $\alpha_{t,i}$ represents the unit response volume incented by the reward price. *m* represents the scale factor between the unit response volume and the electricity consumption. *n* represents the exponential relationship between the price and the response volume. The different response behavior of the consumers can be simulated by adjusting the set value of *m, n,* $\alpha_{t,i}$ and $th_{t,i}$.

Data generation of demand response behavior

A large amount of data is required in training the models of DQN and SOMs. Sufficient key data is the basic condition of the proposed method. Literatures [22–26] use official databases, survey sampling and other methods to collect the data required for the model, which provide a good reference for related research that requires a large amount of data. Similarly, this paper uses the demand response price data provided by the official website of the Henan Provincial Development and Reform Commission, and combines demand response model established with Eq. (1) to generate the demand response behavior data required for model training.

(1) Simulation of response characteristics of different consumers

Due to the factors such as lifestyle and work arrangements, the electricity consumption profile of the consumers varies during the day. Concisely but generally, this paper sets $m = 12$, $n = 1$, which means the response volume of consumer $res_{t,i}$ is linearly correlated with the reward price $th_{t,i}$.

Because of the factors such as one's own economic level and lifestyle, each class of the consumers expect different reward price $th_{t,i}$ and provide different response volume by the reward price $\alpha_{t,i}$. Since there is no large-scale open access data set of small and medium-sized consumers participated in demand response market, this paper uses the Monte Carlo simulation method to generate response behavior data of 10,000 consumers, according to the demand response model of consumer described in Eq. (1) and the actual electricity consumption patterns of small and medium-sized consumers [27].

Table 1
The range of bidding volume and bidding price of the aggregator.

Period	Bidding volume (MW·h)	Price(¥/kW·h)
t_1	[−22,−18]	[12,52,13.52]
t_2	[20,24]	[13.00,14.08]
t_3	[30,36]	[16.00,18.26]
t_4	[24,28]	[14.08,15.32]
t_5	[28,32]	[15.32,16.72]
t_6	[−24,−20]	[13.00,14.08]

The parameters $\alpha_{t,i}$ and $th_{t,i}$ of the consumers are set to obey normal distribution, and the mean and variance of the normal distribution corresponding to different class of consumers are different, which reflects the different response characteristics. The adopted normal distribution is shown in Eq. (2).

$$\begin{cases} f\left(\alpha_{t,i}\right) = \frac{1}{\sqrt{2\pi}\sigma_{\alpha,i}} \exp\left[-\frac{\left(\alpha_{t,i}-\mu_{\alpha,i}\right)^2}{2\sigma_{\alpha,i}^2}\right] \\ f\left(th_{t,i}\right) = \frac{1}{\sqrt{2\pi}\sigma_{th,i}} \exp\left[-\frac{\left(th_{t,i}-\mu_{th,i}\right)^2}{2\sigma_{th,i}^2}\right] \end{cases} \tag{2}$$

where μ represents expectation and σ represents standard deviation.

(2) Division of response periods

In order to show the mechanism of the proposed method in a compact way, this paper divides a day into 6 periods, and the response pattern of each consumer differs in these 6 periods, therefore the solving speed and the diversity of response patterns can be balanced.

To better reflect the real response patterns of the consumers, this paper assumes that the price elasticity of a group of consumers which is composed of 500 consumers varies periodically, that is, the peak value of price elasticity sequentially appears at the 6 periods.

Decision-making behavior of the aggregator

The bidding of the aggregator in the wholesale market

In the wholesale market, if the bid volume of the aggregator is D_t, and the bidding price of the aggregator is P_t in period t, then the relationship between D_t and P_t can be expressed by supply function as shown in Eq. (3).

$$P_t = f_s^{-1}(D_t) \tag{3}$$

The reward price of demand response is set to 12–18 |¥/kW· h in this paper, referring to the demand response market rules of Henan Province, China [28]. And the supply function (3) is expressed in detail as Eq. (4).

$$P_t = 10 + 5 \times 10^{-5} \times D_t + 5 \times 10^{-6} \times D_t^2 \tag{4}$$

If the bidding volume D_t of the aggregator meets the transaction requirements and the bidding price P_t doesn't exceed the clearing price P_{sm} of the market, then the aggregator can win the bid. The winning conditions are shown in Eq. (5).

$$\begin{aligned} D_{\min} \leq D_t \leq D_{\max} \\ P_t \leq P_{\text{sm}} \end{aligned} \tag{5}$$

where $D_{\max}$ and $D_{\min}$ are the maximum and minimum limits of bidding volume.

Considering the uncertainty of market transactions, this paper sets the bidding volume of aggregator randomly fluctuate within the feasible range in each period. The corresponding range of bid-winning price shown in Table 1 can be calculated by Eq. (4). The positive value represents load reduction demand response, while the negative value represents load increase demand response.

Decision-making of the aggregator about the consumers

On the other hand, in order to cover his bid-winning volume, the aggregator needs to gain the volume from the consumers whom he represents. Therefore, the aggregator classifies the consumers according to their response patterns, and then sends appropriate reward price to each class of consumer.

The objective of this decision-making is to maximize the profit R_t of the aggregator when he trades with the consumers during period t, considering that the bidding volume in the wholesale market can be delivered. The objective function is shown in Eq. (6).

$$\max R_t = \sum_{i=1}^{k} R_{t,i} = (P_t - p_{xt}^{sm}) \sum_{i=1}^{k} res_{t,i}$$
$$\sum_{i=1}^{k} res_{t,i} \geq D_t \tag{6}$$

where $R_{t,i}$ represents the profit of the aggregator when the *i* th class consumers to be selected to participate in demand response. *p*sm *xt* represents the marginal reward price. $res_{t,i}$ represents the response volume of the *i* th class consumers during period *t*. Assuming that the *N* consumers agented by the aggregator are classified into *K* classes, and a total response volume from *k* classes of consumer can meet the bid-winning volume D_t.

By substituting Eq. (1) into Eq. (6), the profits of the aggregator when trading with the *i* th class consumers can be obtained, which is shown in Eq. (7).

$$R_{t,i} = \alpha_{t,i}(P_t - p_{xt,i})(p_{xt,i} - th_{t,i})^n \tag{7}$$

By using the substitution method, make $u = p_{xt,i} - th_{t,i}$, i.e.$p_{xt,i} = u + th_{t,i}$, so Eq. (7) can be written as:

$$R_{t,i} = \alpha_{t,i}(P_t - th_{t,i} - u)u^n \tag{8}$$

Furtherly transformed, Eq. (8) is obviously a univariate *n + 1*° function of *u*, specifically:

$$R_{t,i} = -\alpha_{t,i}u^{n+1} + \alpha_{t,i}(P_t - th_{t,i})u^n \tag{9}$$

Deriving Eq. (9), the derivation of the aggregator's profit can be obtained, which is shown in Eq. (10).

$$R'_{t,i} = -(n+1)\alpha_{t,i}u^n + n\alpha_{t,i}(P_t - th_{t,i})u^{n-1} \tag{10}$$

Let the derivation equals to 0, it can be obtained that $u = \frac{n(P_t - th_{t,i})}{n+1}$, i.e. $p_{xt,i} - th_{t,i} = \frac{n(P_t - th_{t,i})}{n+1}$. Then we can obtain $p_{xt,i} = \frac{th_{t,i} + nP_t}{n+1}$. Considering Eq. (10), it can be seen that the derivations are with reverse signs when the value of the independent variable $p_{xt,i}$ is on the left and right side of $\frac{th_{t,i} + nP_t}{n+1}$ respectively. And when $P_t - th_{t,i}$ is greater than 0, the derivation on the left side of $p_{xt,i}$ is greater than 0, while the derivation on the right side of $p_{xt,i}$ is less than 0.

Therefore, $p_{xt,i} = \frac{th_{t,i} + nP_t}{n+1}$ is the maximum of Eq. (7) when $P_t - th_{t,i}$ is greater than 0, in other words, the market price for the aggregator in the wholesale market is greater than the expected reward price of the consumer.

In summary, it can be concluded that the optimal reward price *p*max *xt,i* and the maximum profit *R*max *t,*of the aggregator in the transaction with *i* th class consumers are shown in Eq. (11) and (12) respectively.

$$p_{xt,i}^{\max} = \frac{th_{t,i} + nP_t}{n+1} \tag{11}$$

$$R_{t,i}^{\max} = \alpha_{t,i}\frac{n^n}{(n+1)^{n+1}}(P_t - th_{t,i})^{n+1} \tag{12}$$

On the other hand, if the aggregator has already sent the optimal reward price to all the classes of consumers, and the response volume from all the consumers still cannot meet the bid-winning volume, then the aggregator needs to increase the reward price. When the marginal reward price *psm xt* continually increases, and finally equals to the market price, the profit of the aggregator in this transaction is 0. Therefore, it is necessary to impose certain restrictions on the theoretical maximum bidding volume *Dmax t*of the aggregator, in order to ensure that the profit of aggregator is not negative, which is shown in Eq. (13).

$$D_t^{\max} = \sum_{i=1}^{K} \alpha_{t,i}(P_t - th_{t,i})^n \tag{13}$$

Dynamic classification of consumer response behavior based on SOMs

Accurately perceiving the response behavior and acceptable reward price of the consumers is important for maximizing profit of the aggregator. Given the number of consumers, classification of consumers must be done. The demand price elasticity reflects the price sensitivity of the consumers to commodity demand, which can reflect the demand response ability of the consumers [29].

Affected by living habit, work arrangement and income in real life, the electricity consumption patterns of consumers are different. And the electricity consumption patterns of consumers may change periodically due to the cyclical factors, such as season and work schedule. When the electricity consumption pattern of consumers changes, if reward price is still pushed according to the original classification, consumers will be unable to provide the expected demand response volume corresponding to the reward price. Therefore, it is necessary to dynamically classify consumers to accurately obtain the actual electricity consumption patterns of consumers in real time. SOMs is a kind of neural network that performs a projection of high-dimensional data set from the original input space to the two-dimensional output space [9]. SOMs algorithm has unique advantages in the process of high-dimensional data, complex data distribution, and result visualization. It can clearly present the distribution of consumer classification results, which is very suitable for the consumer classification proposed in this paper. Combining the training process of SOMs algorithm with the network training of DQN algorithm and dynamically classify consumers by SOMs algorithm based on the updated data from the experience pool can effectively reflect the changes in the electricity consumption patterns of consumer. So, the decision errors caused by classification can be reduced. Therefore, SOMs algorithm is proposed to classify consumers dynamically.

Step 1 Initialization of the weight vector in output layer

The input layer of SOMs consists of N nodes, representing N consumers. The number of nodes in core layer is set to M, the weight vector connecting node i in output layer and node j in core layer is represented as W_{ij}。

Firstly, assign small random number to each weight vector in output layer and perform normalization processing, obtaining W_{ij}. And then establish initial winning neighborhood $S^*(0)$ and initial value of earning rate η.

Step 2 Input and normalization of the consumer price elasticity

The equivalent electricity price $p' t$ for the consumers participating in demand response during the period t is shown in Eq. (14). And the consumer elasticity ε_t can be derived from Eq. (14), which is shown in Eq. (15).

$$p'_t = \frac{p_t q'_t - p_{xt}(q_t - q'_t)}{q'_t} \tag{14}$$

$$\varepsilon_t = \frac{\Delta q_t}{q_t}\frac{p_t}{\Delta p_t} = \frac{q'_t - q_t}{q_t}\frac{p_t}{p'_t - p_t} = \frac{p_t q'_t}{p_{xt} q_t} \tag{15}$$

where p_t is the original electricity price in period t. p_{xt} is the average reward price, q_t is the baseline electricity consumption of consumer, and $q' t$ is the average electricity consumption of consumer after demand response. All the above data are available historical data.

The aggregator can continuously calculate and update the electricity demand price elasticity of the consumers according to Eq. (15). The self-elasticity matrix E_i of i th class consumer in different periods is shown in Eq. (16), which provides the data basis for classification of consumer response ability.

$$\boldsymbol{E}_i = \left[\varepsilon_{1,i}, \varepsilon_{2,i}, \varepsilon_{3,i}, \varepsilon_{4,i}, \varepsilon_{5,i}, \varepsilon_{6,i}\right] \tag{16}$$

After normalization, the consumer self-elasticity matrix E_i can be used as input data for the SOMs.

Step 3 Searching for winning node

According to competitive learning rule, each input node can only activate one neuron in core layer at the same time, which is called winning node. The winning node corresponding to consumer i can be calculated by the similarity between nodes in input layer and nodes in core layer measured by Euclidean distance. The calculation method is shown in Eq. (17).

$$k = \arg\max_j ||W_{ij}^{\mathrm{T}} E_i|| \tag{17}$$

Step 4 Adjustment of weight vector in the winning neighborhood

The winning neighborhood $S(k)$ is a weight adjustment field determined with the winning node k as the center. The radius of the winning neighborhood $S(k)$ will be assigned a larger initial value, and then gradually contracts during training. The weight vectors within the winning neighborhood S (k) will be adjusted according to the input vector. The adjustment method is shown in Eq. (18).

$$W_{ij} = W_{ij} + \eta(E_i - W_{ij}), W_{ij} \in S(k) \tag{18}$$

where $\eta\in[0,1]$ is the learning rate, which gradually decreases during training.

Step 5 Determination of learning rate

When the learning rate η is less than the set value $\eta_{\min}$, the training will end. At this point, each node in the input layer will establish a unique and determined mapping relationship with a node in the core layer, and the input vectors corresponding to adjacent nodes in the core layer will be considered as the same type, then the classification of consumers based on elasticity characteristics can be achieved.

If the end condition is not met, return to step 2 to continue training.

Decision-making of the aggregator based on deep reinforcement learning

The agented consumers are numerous and have various electricity consumption behaviors, so it is difficult to build the model (as shown in Eq. (1)) which contains the key parameters of response behavior, and support analysis of demand response based on the complete information model. Because the complete information model means the aggregator can accurately obtain the response behavior parameter of consumers in the lower transaction, and the bi-level transaction model can be solved by conventional optimization method. From a mathematical perspective, conventional optimization method is effective only when all parameters of the mathematical programming model are known.

Alternatively, with means of the model-free algorithm of deep reinforcement learning, the aggregator is capable of optimal decision-making through training with extensive historical experience. Therefore, this paper uses deep reinforcement learning to realize the consumer behavior perception and consumer demand response decision-making for the aggregator. The demand response transaction between the aggregator with the consumers is described as the interaction between environment and agent. The profit of the aggregator can be regarded as the reward. The decision-making ability of agent continuously increases through the training. Finally, the aggregator can achieve maximum profit by accurately selecting the suitable class of response consumer and incenting them with the appropriate reward price.

Deep Q network (DQN) is used to solve decision-making that maximizes the transaction profit of the aggregator, including reasonable selection of response consumers and modification of reward price. Asynchronous offline learning mode is adopted in DQN, and experience replay mechanism is used to training neural network. The optimal strategy based on the Q value is evaluated and improved through ε-*greedy* strategy.

Reward and action setting of the aggregator

After winning the bid in wholesale market, the aggregator needs to deliver the bid-winning volume of demand response, while maximizes his profit. Therefore, the aggregator needs to select promising consumers to participate in demand response from numerous consumers in a local transaction, and reward the selected consumers.

Action setting of the aggregator

Summarizing the demand response transaction between the aggregator and the consumers, the action sequence of the aggregator is as follows: (1) Select the i th class consumers to participate in demand response from the K-class consumers classified by SOMs; (2) Determine the optimal reward price $p_{xt,i}$ for the selected i th class consumer.

The action is shown in Eq. (19).

$$\begin{aligned} a &= \{a_1, a_2\} \\ a_1 &= i, i \in [1, K] \\ a_2 &= \{p_{xt,i} - 1, p_{xt,i}, p_{xt,i} + 1\} \end{aligned} \tag{19}$$

where a_1 and a_2 represent the selection of the consumers and the determination of the optimal reward price by the aggregator. i represents the class of the selected consumers. K represents all the classes of the consumers.

The aggregator executes actions, selects the i th class consumers and sends the reward price $p_{xt,i}$ to the consumers, and then obtains the possible response volume $res^*_{t,i}$ from the consumers. In the j-th interaction, if the undelivered bid-winning volume $D^{\text{obs}}_{t,j} > 0$, the aggregator needs to continue executing action, guides consumers to offer more demand response volume $res^*_{t,i}$, and updates the remaining undelivered bid-winning $D^{\text{obs}}_{t,j+1}$, which is shown in Eq. (20).

$$D^{\text{obs}}_{t,j+1} = D^{\text{obs}}_{t,j} - res^*_{t,i} \tag{20}$$

The termination condition of aggregator action is the remaining undelivered volume $D^{\text{obs}}_{t,j+1} \le 0$.

Reward setting of the aggregator

When the aggregator selects i th class consumers and sends reward price $p_{xt,i}$ to the consumers, the consumers will provide response volume $res^*_{t,i}$. The action reward of aggregator is defined as Eq. (21).

$$r_{t,i} = \begin{cases} -(p_t - p_{xt,i}), res^*_{t,i} = 0 \\ 0, res^*_{t,i} > 0 \& D^{\text{obs}}_{t,j+1} > 0 \end{cases} \tag{21}$$

When the response volume of the consumers $res^*_{t,i}$ is 0, the artificially set penalty term will suggest the aggregator that this class of consumers should not be selected for this condition. When the response volume of the consumers $res^*_{t,i}$ is greater than 0, but the undelivered bid-winning $D^{\text{obs}}_{t,j+1}$ still remains, the action reward of the aggregator is set to 0, making the aggregator continue executing actions.

When the undelivered volume $D^{\text{obs}}_{t,j+1}$ is 0, i.e. the aggregator has delivered all the bid-winning volume of demand response, the aggregator settles with all consumers participating in the demand response transaction based on the clearing price, and obtains the accumulated reward R_t in this transaction, which is shown in Eq. (22).

$$R_t = (p_t - p^{sm}_{xt}) \sum_{i=1}^{k} res^*_{t,i} + \sum_{i=1}^{k} r_{t,i} \tag{22}$$

If the aggregator has rewarded all the agented consumers in the local transaction, but the gained response volume from the consumers still cannot meet the bid-winning volume that needed to be delivered, then the aggregator needs to increase reward price for the consumers, along with a loss of profit, which is shown in Eq. (23).

$$R_t = -\sum_{i=1}^{k} p_{xt,i} res^*_{t,i} + \sum_{i=1}^{k} r_{t,i} \tag{23}$$

Determining the optimal reward price

This section will discuss the action a_2—the decision-making about the optimal reward price that the aggregator sends to the consumers. From Eq. (1), it can be seen that there must exist an optimal reward price $P^{\max}_{xt,}$ within the price range $[th_{t,i,}\ P_t]$ for any class i of consumer, from perspective of maximizing the profit $R_{t,i}$ of the aggregator. However, the response behavior parameter of consumers is hard to accurately obtained, so the optimal reward price $P^{\max}_{xt,}$ can not be gotten by solving Eq. (11) directly.

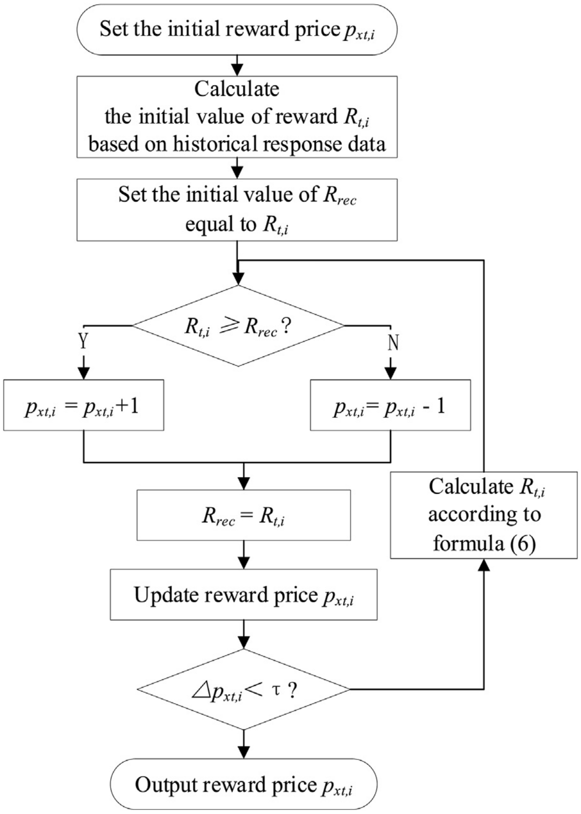

Set the initial reward price $p_{xt,i}$
Calculate the initial value of reward $R_{t,i}$ based on historical response data
Set the initial value of R_{rec} equal to $R_{t,i}$
$R_{t,i} \geqslant R_{rec}$?
Y
N
$p_{xt,i} = p_{xt,i}+1$
$p_{xt,i}= p_{xt,i}$ - 1
$R_{rec} = R_{t,i}$
Calculate $R_{t,i}$ according to formula (6)
Update reward price $p_{xt,i}$
$\triangle p_{xt,i} < \tau$?
Output reward price $p_{xt,i}$

Fig. 2. Flow chart of the method for determining the optimal reward price.

This paper proposes the method of determining the optimal reward price $P^{\max}_{xt,}$ based on the “trial and error method” learning process of DQN algorithm, which is shown in Fig. 2.

First, the reward price $p_{xt,i}$ is set at a small value as the initial value. When the *i* th class consumers is selected for the first time in period *t*, the aggregator calculates the reward $R_{t,i}$ corresponding with the initial reward price $p_{xt,i}$ on the basis of the historical response data. And the value of $R_{t,i}$ will be assigned to R_{rec}. And then the reward price $p_{xt,i}$ will be increased by 1 unit.

When the *i* th class consumers is selected in next iteration, the aggregator calculates the reward $R_{t,i}$ corresponding with the present reward price $p_{xt,i}$, and compares the values of $R_{t,i}$ with R_{rec}. If $R_{t,i}$ is greater than R_{rec}, which proves that the action to increase reward price in this iteration is unreasonable, then the reward price $p_{xt,i}$ will be reduced by 1 unit. Conversely, the reward price $p_{xt,i}$ will be increased by 1 unit.

Until the variation of reward price is less than the tolerance error τ, the optimal reward price *pmax xt,i* can be determined.

Action selection of the aggregator based on ε-greedy strategy

In the proposed method, the action selection of the agent representing the aggregator adopts the *ε-greedy* strategy.

The aggregator will choose the action with the highest in each round with a high probability *1-ε*, while the actions with other Q value are selected with a smaller probability ε to avoid trapping in the previous strategy, so that DQN algorithm may try other strategies with lower estimated value but better actual performance. The *ε-greedy* strategy is shown in Eq. (24).

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A(s)|}, a = \arg\max\limits_{a} Q(s,a) \\ \frac{\varepsilon}{|A(s)|}, a \neq \arg\max\limits_{a} Q(s,a) \end{cases} \tag{24}$$

Deep learning about decision-making ability of the aggregator

In the interaction between the agent and environment, each iteration will generate an experience e_{xp}, as shown in Eq. (25). Through the continuous accumulation of experience, a limited and constantly updated experience pool can be structured. In an offline mode, the deep Q network selects samples from the experience pool through the randomly sampled experience replay mechanism and conducts deep learning.

$$e_{xp} = \left\{ s, a, r, s' \right\} \tag{25}$$

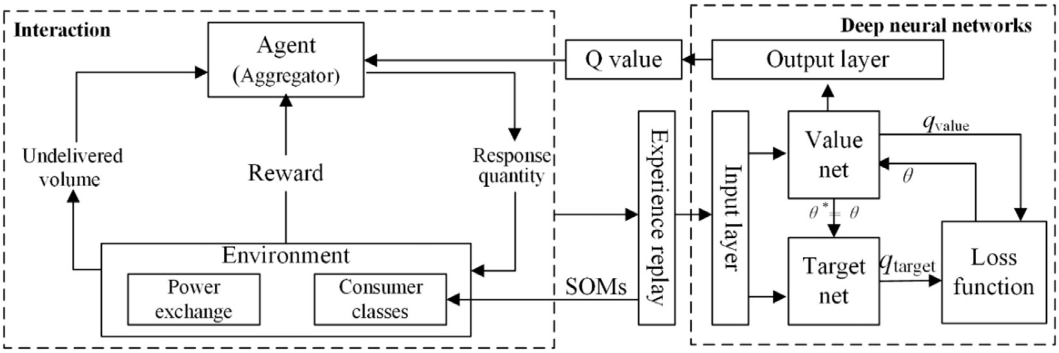


Fig. 3. Demand response perception and decision-making based on deep reinforcement learning.

where s represents the current state of the iteration, i.e. the undelivered volume $D_{t,i}^{\text{obs}}$. a represents the action of the agent toward state s, i.e. the aggregator selects the i th class consumers to participate in demand response transaction. r represents the reward for executing the action, i.e. $r_{t,i}$ represents the action reward of the aggregator in Eq. (21). s' represents the state of the next iteration, i.e. the remaining undelivered response volume $D_{t,i+1}^{\text{obs}}$.

The action-value function under a certain strategy is fitted in deep reinforcement learning by deep neural network. As the direct basis for action selecting, the Q value outputted by deep neural network will determine the selection sequence of responsive consumer and affect the profit of the aggregator.

This paper uses offline mode for learning to train value neural network and target neural network. The value network is designed to calculate the current value of action-value function with parameter θ. The target network is designed to calculate the predicted value of action-value function with another parameter θ^*, and it is updated at a low frequency.

During training, an experience is randomly selected from the experience pool, then the corresponding parameters s and a are put into the value network and target network, resulting in q_{value} and q_{target}. According to gradient descent method, the value of loss function is continuously reduced during training for searching parameter θ of the value network. The target network parameter is updated to $\theta^*=\theta$ after a preset number of learning times.

The output of target network q_{target} is shown in Eq. (26), and the definition of the loss function L is shown in Eq. (27).

$$q_{\text{target}} = r + \gamma \max_{a'} Q(s', a', \theta) \tag{26}$$

$$L = [q_{\text{target}} - q_{\text{value}}]^2 \tag{27}$$

The above process can be represented by the interaction between the intelligent agent (aggregator) and environment as shown in Fig. 3.

Simulation result and discussion

Effectiveness verification of classification based on SOMs

SOMs is used to classify the consumers with different electricity price elasticity. The effectiveness of classification based on SOMs needs to be fully verified. Literature [30] fully demonstrates the discussion from the aspects of changes in the situation of the same object over time and comparison of results between different objects. The discussion of this paper refers to the above literature, the effectiveness verification of classification based on SOMs is organized from the two aspects of the changes in consumer classification results over time and classification results among different consumers.

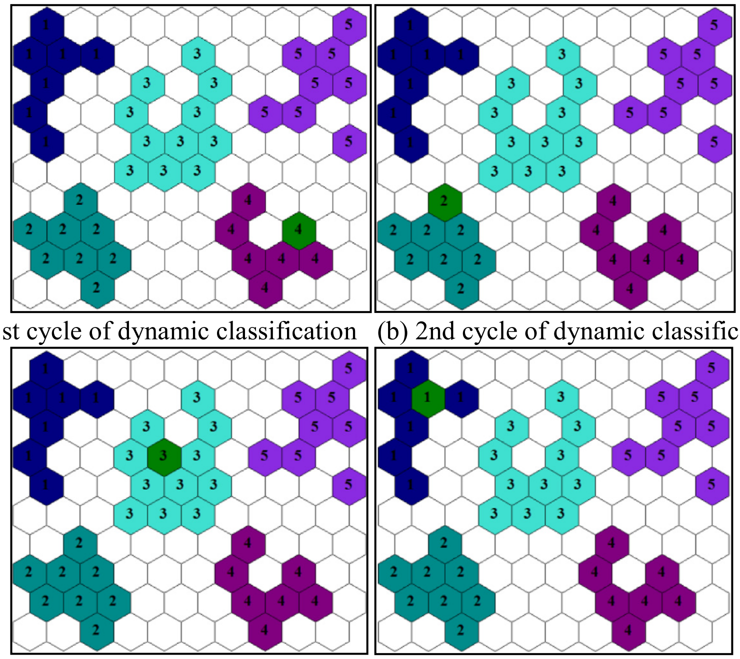
The results of the dynamic consumer classification are shown in Fig. 4.

The distribution of winning neuron in the competitive layer is shown in Fig. 4. The meshes with the same number and color represent a category of consumers. The boundaries between each category are clear and the distribution of each category is relatively compact, which proves that the classification is effective. The dark green grid in Fig. 4 represents 100 consumers with time-varying electricity consumption patterns. In the process of dynamic classification, this category of consumers is classified to category 4, category 2, category 3 and category 1 respectively. This result is consistent with the preset condition in which these consumers are set a change pattern. In summary, the effectiveness of the dynamic classification of SOMs can be fully proven.

The classification results of different consumers are shown in Fig. 5.

From Fig. 5, it can be seen that all the demand response consumers are classified into 5 categories, each has its own price elasticity temporal distribution characteristics.

The appearing times of their peak price elasticity are: 1st-class consumers in the period of 16:00 to 20:00, the 2nd-class of consumers in the period of 8:00 to 12:00, the 3rd-class of consumers in the period 12:00 to 16:00, the 4th-class of consumers in the period of 16:00 to 20:00, and the 5th-class of consumers in the period of 20:00 to 24:00. The consumer classification results of SOMs conform to the preset single-peak pattern, thus the consumer classification is effective.



(a) 1st cycle of dynamic classification (b) 2nd cycle of dynamic classification

(c) 3rd cycle of dynamic classification (d) 4th cycle of dynamic classification

Fig. 4. The change of dynamic classification over time.

Analysis on learning effectiveness of DQN

The value change of loss function during training of deep neural networks is shown in Fig. 6.

From Fig. 6, it can be seen that the value of loss function decreases rapidly with the increase of training times, and converges to a small value at the 500th training. In the subsequent training, the value of loss function still decreases and gradually converges to an even lower value. It can be concluded that the evaluation value of value network is ultimately close to the predicted value of target network, and the training of deep neural network is stable and convergent, can result in effective deep learning.

Analysis on consumer selection and reward price decision of DQN

The simulation in this paper takes the decision-making of consumer selection based on the complete information model (based on complete information and mathematical programming) as contrast group. In every 2000 training times, the decision results of the agent based on DQN algorithm are compared with the results of the contrast group. The error rate is defined as the degree of difference between the two methods. The results are shown in Fig. 7.

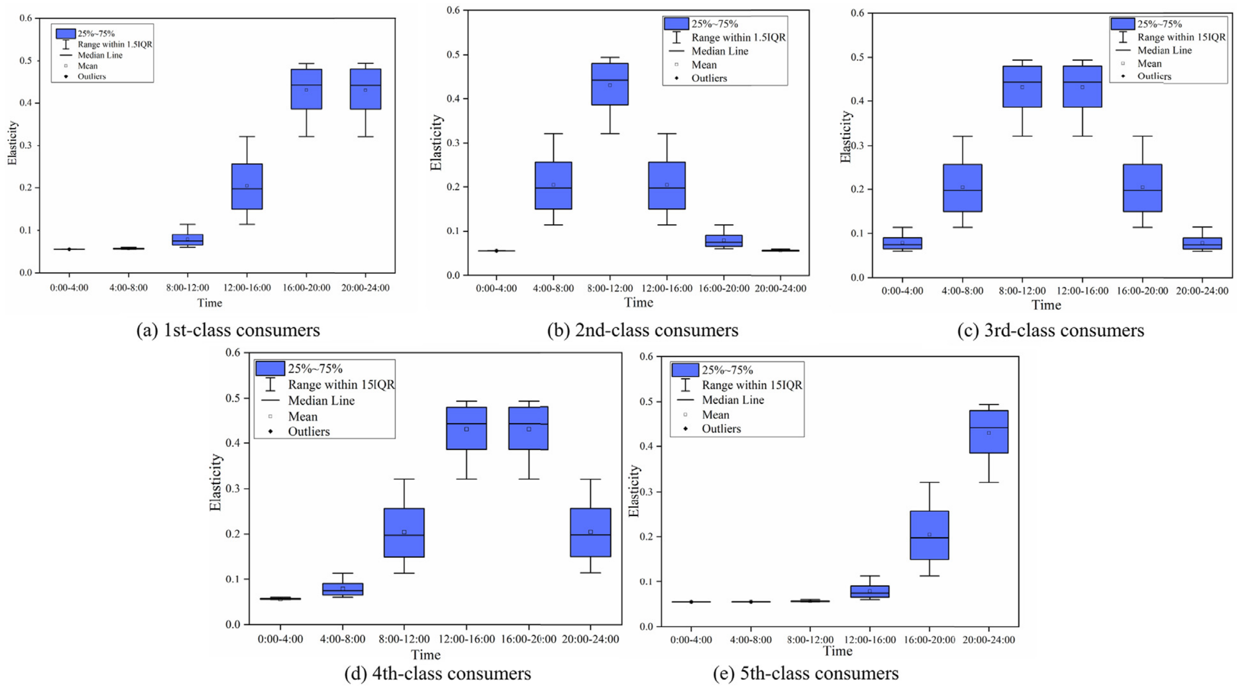
From Fig. 7, it can be seen that as the times of training increases, the error rate of consumer selection ability based on DQN algorithm gradually decreases with different shape in each period. Finally, the error rate of each period tends to less than 10 %, indicating that the agent works well in accurate consumer selection.

Fig. 8 shows the comparison of the response reward prices of 1st-class consumers solved by DQN and complete information model separately.

To conform to objective laws better, the electricity consumption of consumer is assumed to fluctuate randomly within a range, thus the reward price solved by complete information model is expressed in an interval. For computational facilitation, the minimum modification of the reward price by the agent is set to 1 unit, thus the results are all integers. And as the electricity consumption by consumers fluctuates, the reward price also fluctuates.

From Fig. 8, it can be seen that the reward prices solved by DQN are basically within the range of prices solved by the complete information model, which indicates that the agent has produced the reasonable reward price.

If the consumer model parameters in Eq. (6) are known, the transaction strategy of the aggregator in the local transaction can be directly solved by the Eq.. However, consumer information is difficult to be fully acquired. Based on deep reinforcement learning about the historical data, the agent can make the demand response trading profits of the aggregator increase and gradually approach to the theoretical maximum. Fig. 9 shows the unit rewards in demand response transaction of DQN method and the complete information model. In order to avoid the influence of excessive numerical fluctuations to observation, each point on the curve in Fig. 9 represents the average of unit reward in every 100 transactions. It can be seen that with the increase of training times, the reward of the agent for each decision gradually increases. When the training times reach more than 7000, the unit reward of the proposed method tends



(a) 1st-class consumers (b) 2nd-class consumers (c) 3rd-class consumers

(d) 4th-class consumers (e) 5th-class consumers

Fig. 5. Self-elasticity distribution of consumers classified by SOMs.

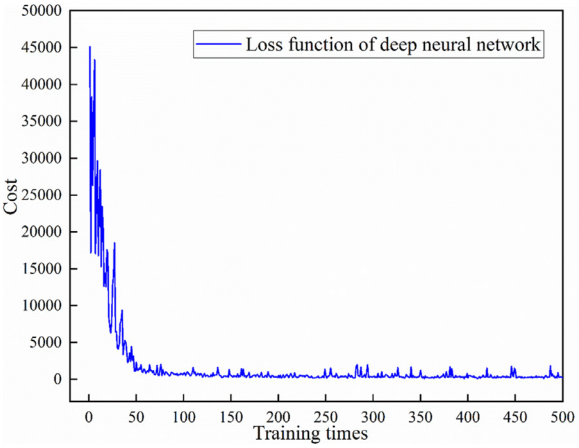


Fig. 6. Value change of loss function of deep neural network.

to be stable and close to the unit reward of complete information model, which proves that the proposed method can effectively realize the perception and decision of demand response.

After training, the unit reward of the method proposed in this paper is still slightly lower than the reward based on complete information model. The reasons are as follows: 1) ε-*greedy* strategy is used in the action selection of DQN. When ε reaches the maximum value of 0.95 in the training, other actions can still be selected with the probability of 5 %. 2) A penalty term for the inaction consumers who have been selected by the aggregator in the local transaction is defined in Eq. (21), which aims at making the agent better learn the selecting sequence of consumers. However, it doesn't actually be executed by the aggregator in the transaction.

Limitations

The method proposed in this paper is mainly based on DQN algorithm and SOMs algorithm. A large amount of historical data is required by the training of model. Therefore, for some scenarios that demand response market has not been developed or the demand response behavior data of consumers is lack, the method proposed in this paper may perform poorly.

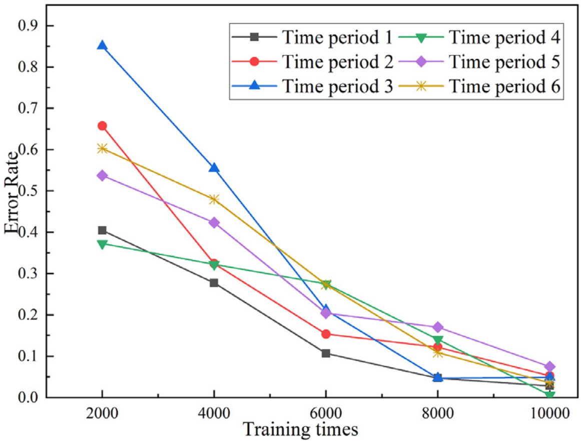


Fig. 7. Error rate of DQN result compared to the result of complete information model.

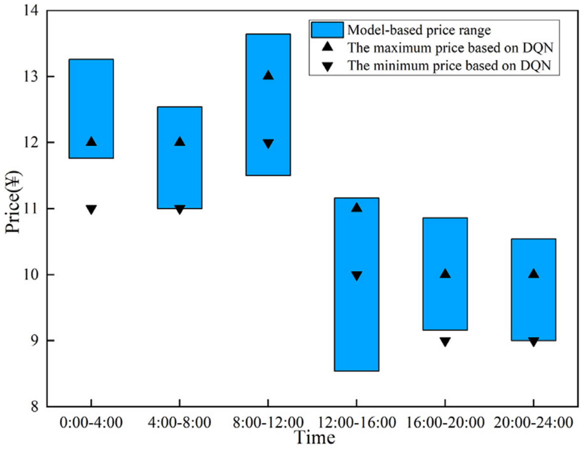


Fig. 8. Comparison of reward prices for the 1st-class of consumers of different algorithms.

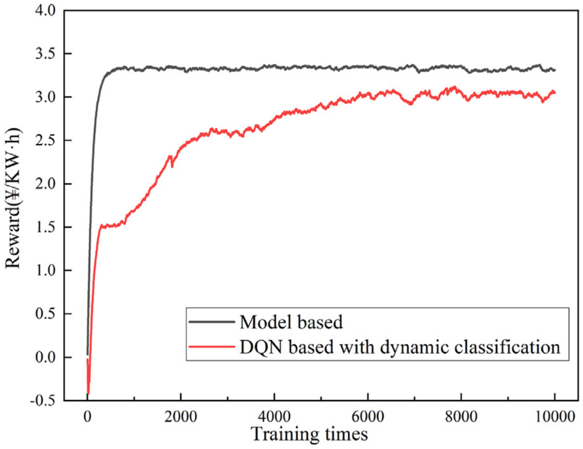


Fig. 9. The aggregator rewards of the complete information model-based algorithm and the DQN algorithm.

Conclusion

In the case of an aggregator agents for many small and medium-sized consumers to participate in wholesale demand response market, there are problems of inaccuracy prediction of the response volume and difficulty of timely trading decision making. To solve the problems, this paper proposes a method of demand response perception and decision-making based on dynamic classification combining DQN and SOMs. The analysis and simulation results indicate that:

With the complete information of consumers, the aggregator can optimize trading strategy to maximize profit based on given model, determine the selection sequence, optimal incentive, and bidding volume of the responsive consumers.

Under the condition of incomplete trading environment information, the agent based on deep reinforcement learning algorithm can learn response pattern of the consumers from historical trading experience, and combine them with the state of the market to make trading decisions on response consumer selecting and reward price modifying. When the consumer information can be continuously updated, the dynamic classification of consumers based on SOMs algorithm can effectively restrict prediction errors caused by uncertainty of response behavior, and improve the perception ability and decision-making efficiency of the aggregator. On the premise of meeting the real-time trading requirements, the proposed method can effectively control the deviation between the actual response volume of the consumers in the local transaction and the bid-winning volume in the wholesale market, and enable the aggregator to make a satisfactory profit towards the theoretical optimal value.

It is found that accurate consumer information is very important for the demand response decision-making of the aggregator. Therefore, relevant policies and mechanisms should be Eq.ted by the government and the market regulator which can guide consumers to provide necessary information, and cooperate with the aggregators to facilitate demand response. At the same time, specifications regarding the use of consumer information should be developed to ensure the information security of the consumer.

Practically, the electricity consumption adjustment of consumers is time coupling, that is, he may shift his consumption from one period to another period. This mutual elasticity is not considered in the paper, and will be further studied in the future.

Ethics statements

Our work complies with the ethical guidelines required by MethodsX.

And our work does not relate to human subjects, animal experiments and data collected from social media platforms.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Guangda Xu: Conceptualization, Methodology, Formal analysis, Software. **Shihang Song:** Writing – original draft, Data curation, Visualization. **Yu Li:** Validation, Project administration. **Yi Lu:** Data curation, Funding acquisition. **Yuan Zhao:** Supervision. **Li Zhang:** Writing – review & editing. **Fukun Wang:** Visualization. **Zhiyu Song:** Visualization.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the science and technology project of North China Electric Power Research Institute, which is “Research on key technologies for power quality evaluation and improvement of new distribution network based on collaborative interaction of source-network-load-storage” (KJZ2022016).

Supplementary material *and/or* additional information [OPTIONAL]

N/A

References

[1] B. Parrish, P. Heptonstall, R. Gross, A systematic review of motivations, enablers and barriers for consumer engagement with residential demand response, Energy Policy 138 (2020) 1–11.
[2] N.I. Nwulu, X.H. Xia, Optimal dispatch for a microgrid incorporating renewables and demand response, Renew. Energy 101 (2017) 16–28.
[3] A.H. Mohsenian-rad, A. Leon-garcia, Optimal residential load control with price prediction in real-time electricity pricing environment, IEEE Trans. Smart. Grid. 1 (2) (2010) 120–133.
[4] S. Nan, M. Zhou, G. Li, Optimal residential community demand response scheduling in smart grid, Appl. Energy 210 (2018) 1280–1289.
[5] D. Setlhaolo, X. Xia, J. Zhang, Optimal scheduling of household appliances for demand response, Electr. Power Syst. Res. 116 (2014) 24–28.
[6] M. Shafie-Khah, P. Siano, A stochastic home energy management system considering satisfaction cost and response fatigue, IEEE Trans. Industr. Inform. 14 (2) (2017) 629–638.
[7] D. Zhang, L.I. Shuhui, M. Sun, An optimal and learning-based demand response and home energy management system, IEEE Trans. Smart. Grid. 4 (7) (2016) 1790–1801.
[8] M. Sun, T. Zhang, Y. Wang, Using Bayesian deep learning to capture uncertainty for residential net load forecasting, IEEE Trans. Power Syst. 35 (1) (2020) 188–201.

[9] S. Valero, M. Ortiz, C. Senabre, Methods for customer and demand response policies selection in new electricity markets, IET Gener. Trans. Distrib. 1 (1) (2007) 1.

[10] M. Rahimiyan, H.R. Mashhadi, An adaptive-learning algorithm developed for agent-based computational modeling of electricity market, IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 40 (5) (2010) 547–556.

[11] T. Chen, W.C. Su, Local energy trading behavior modeling with deep reinforcement learning, IEEE Access. 6 (2018) 62806–62814.

[12] T. Chen, W.C. Su, Indirect customer-to-customer energy trading with reinforcement learning, IEEE Trans. Smart. Grid. 10 (4) (2018) 4338–4348.

[13] H.W. Wang, T.W. Huang, X.F. Liao, H. Abu-Rub, G. Chen, Reinforcement learning for constrained energy trading games with incomplete information, IEEE Trans. Cybern. 47 (10) (2017) 3404–3416.

[14] M. Volodymyr, K. Koray, S. David, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

[15] D. Silver, J. Schrittwieser, K. Simonyan, Mastering the game of Go without human knowledge, Nature 550 (7676) (2017) 354–359.

[16] H.C. Xu, H.B. Sun, D. Nikovski, S. Kitamura, K. Mori, H. Hashimoto, Deep reinforcement learning for joint bidding and pricing of load serving entity, IEEE Trans. Smart. Grid. 10 (6) (2019) 6366–6375.

[17] X. Huang, S.H. Hong, M. Yu, Y. Ding, J. Jiang, Demand response management for industrial facilities: a deep reinforcement learning approach, IEEE Access. 7 (2019) 82194–82205.

[18] B. Wang, Y. Li, W. Ming, S. Wang, Deep reinforcement learning method for demand response management of interruptible load, IEEE Trans. Smart. Grid. 11 (2020) 3146–3155.

[19] R. Lu, S.H. Hong, Incentive-based demand response for smart grid with reinforcement learning and deep neural network, Appl. Energy 236 (2019) 937–949.

[20] Y. Liu, C. Yang, L. Jiang, Intelligent edge computing for iot-based energy management in smart cities, IEEE Netw. 33 (2) (2019) 111–117.

[21] M.H. Albadi, E.F. El-Saadany, A summary of demand response in electricity markets, Elect Power Syst. Res. 78 (11) (2008) 1989–1996.

[22] Q. Meng, Z. Yan, A. Shankar, M. Subramanian, Human-computer interaction and digital literacy promote educational learning in pre-school children: mediating role of psychological resilience for kids' mental well-being and school readiness, Int. J. Hum. Comput. Interact (2023) 1–15, doi:10.1080/10447318.2023.2248432.

[23] A. Hafeez, W.J. Dangel, S.M. Ostroff, A.G. Kiani, S.D. Glenn, … A.H. Mokdad, The state of health in Pakistan and its provinces and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019, Lancet Glob. Health 11 (2) (2023) e229–e243, doi:10.1016/S2214-109X(22)00497-1.

[24] A.E. Schumacher, H.H. Kyu, A. Aali, C. Abbafati, R. Abbasgholizadeh, … C.J.L. Murray, Global age-sex-specific mortality, life expectancy, and population estimates in 204 countries and territories and 811 subnational locations, 1950-2021, and the impact of the COVID-19 pandemic: a comprehensive demographic analysis for the Global Burden of Disease Study 2021, The Lancet (2024), doi:10.1016/S0140-6736(24)00476-8.

[25] J.D. Steinmetz, K.M. Seeher, N. Schiess, E. Nichols, B. Cao, C. Servili, … T. Dua, Global, regional, and national burden of disorders affecting the nervous system, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021, Lancet Neurol. (2024), doi:10.1016/s1474-4422(24)00038-3.

[26] X. Zhang, M.F. Shahzad, A. Shankar, S. Ercisli, D.C. Dobhal, Association between social media use and students' academic performance through family bonding and collective learning: the moderating role of mental well-being, Educ. Inf. Technol. (Dordr) 28 (14) (2024), doi:10.1007/s10639-023-12407-y.

[27] The Commission for Energy Regulation Group. Electricity smart metering customer behavior trials findings report [EB/OL]. The Commission for Energy Regulation.[2016-08-15].

[28] Notice on the implementation of power demand response in 2019. Development and Reform Commission of Henan Province [EB/OL], 2019.

[29] A. Gabaldón, R. Molina, Residential end-uses disaggregation and demand response evaluation using integral transforms, J. Modern Power Syst. Clean Energy 5 (1) (2017) 91–104.

[30] A.E. Micah, K. Bhangdia, I.E. Cogswell, D. Lasher, B. Lidral-Porter, E.R. Maddison, … J.L. Dieleman, Global investments in pandemic preparedness and COVID-19: development assistance and domestic spending on health between 1990 and 2026, Lancet Glob. Health 11 (3) (2023) e385–e413, doi:10.1016/S2214-109X(23)00007-4.