Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Joint optimization of energy trading and consensus mechanism in blockchain-empowered smart grids: a reinforcement learning approach

Ruohan Wang[1], Yunlong Chen[1], Entang Li[2*], Lixuan Che[3], Hongwei Xin[2], Jing Li[1] and Xueyao Zhang[2]

Abstract

Under the trend of green development, the traditional fossil fuel and centralized energy management models are no longer applicable, and distributed energy systems that can efficiently utilize clean energy have become the key to research in the energy field nowadays. However, there are still many problems in distributed energy trading systems, such as user privacy protection and mutual trust in trading, how to ensure the high quality and reliability of energy services, and how to motivate energy suppliers to participate in trading. To solve these problems, this paper proposes a blockchain-based smart grid system that enables efficient energy trading and consensus optimization, enabling electricity consumers to obtain high-quality, reliable energy services and electricity suppliers to receive rich rewards, and motivating all parties to actively participate in trading to maintain the balance of the system. We propose a reputation value assessment algorithm to evaluate the reputation of electricity suppliers to ensure that electricity consumers receive quality energy services. To minimize the cost, maximize the benefit for the electricity suppliers and optimize the system, we present an algorithm based on reinforcement learning DDPG to determine the power supplier, power generation capacity, and consensus mechanism between nodes to obtain power trading rights in each round. Simulation results show that the proposed energy trading scheme has good performance in terms of rewards.

Keywords Blockchain, Smart grid, Edge computing, Resource allocation, Energy trading

Introduction

Traditional energy industries, such as power companies, were once powered by fully integrated power companies investing in and building transmission and distribution networks [1, 2]. However, due to increasing electrification and energy demand, as well as poor transmission and distribution networks, traditional fossil fuels and centralized utilities are increasingly unable to meet the demand of consumers and suppliers [3]. In addition, under the general trend of clean energy replacing fossil energy and renewable energy replacing non-renewable energy [4, 5], electricity energy trading should develop towards green development [6–9]. In order to build a new electric energy transaction network, which can not only use clean energy efficiently, but also maintain the balance of the energy market by providing a better Quality of Experience (QoE) and maximizing the benefits of suppliers, the distributed energy resources (DERs) has become the focus of current energy research [10, 11].

*Correspondence:
Entang Li
ls_liet@163.com
[1] State Grid Shandong Electric Power Company Marketing Service Center (metrology center), Jinan, Shandong, China
[2] Shandong Luruan Digital Technology CO., LTD, Jinan, Shandong, China
[3] Lixuan Che is with Weifang Vocational College, Weifang, Shandong, China

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

DER system [12] enables the connection between users to form a distributed network, which is a new energy production-supply-consumption system. It is the product of the development of mature new energy technology and energy storage technology [13–16], and the power balance is transferred from the demand side. At present, the energy of large scale DER systems is mainly electric energy [17]. However, the implementation of a DER transaction system is still under study, and the issues of user privacy protection and transaction trust that may be encountered in decentralization need to be addressed. How to provide better QoE and reliable energy services in a transaction, while offering substantial rewards and incentives to suppliers, are among the issues that need to be addressed.

In recent years, blockchain has attracted attention from all walks of life due to its decentralized and tamper proof characteristics. Its essence is a distributed database, which is jointly maintained by all nodes. Blockchain is also used to satisfy the trusted construction of the metaverse [18]. It is the theoretical basis for the implementation of DER transaction systems in [19–21]. With the help of blockchain technology, energy producers and consumers can be directly connected, thus simplifying the mutual relations and interactions between the parties. In [22], the authors compare the electric transaction market based on block chain and existing the difference between electric power market. They point out that the blockchain has broken the boundaries and constraints of the design logic of the contemporary energy market, and are expected to change the traditional centralized energy system through the blockchain. In [23], the author designed a DRE energy transaction authentication mechanism based on blockchain technology applied to distributed energy trading, but this mechanism cannot work in practical production due to its poor throughput.In [24], the author proposes a trading model for energy transaction market to local electric vehicle transactions. Simulation results and evaluation conclude that the blockchain platform improves the autonomy of grid participants but does not involve the overall benefits and rewards of the system network. In [25], the author proposed a heterogeneous computing and resources allocation framework for wireless powered federated edge learning to investigate the performance of the system from users' perspective. They minimize energy consumption and achieve energy harvesting by optimizing the problem. Compared to other methods, this system can achieve efficient federated learning. The overall performance of the power energy trading network is considered and optimized to maximize transmission and minimize consumption, providing a reference for system design. In [26], the authors focus on the cost of individual participants, rather than merely optimizing the cost of the entire process as in existing works. They improved the convergence speed of federated learning by adjusting the local CPU cycle frequency and other related parameters. It can be seen from the experimental results that they have well balanced the cost and fairness. In [27], the author analyzed the transparency of blockchain. Smart contracts make the operational rules of the entire system open and transparent, achieve information symmetry and market effectiveness, and ensure the security and reliability of the trading system. In this paper, we propose a trusted transaction method for blockchain-based manufacturing services.

In [28], the authors propose a blockchain-based approach to manufacturing service composition. The key contributions are the study of dynamic QoS evaluation methods and consensus algorithms, and the design of resource rent-seeking and matching mechanisms based on smart contracts. That approach can adaptively complete the composition of manufacturing services while balancing the privacy, security and openness of transaction information, which greatly enhances the trustworthiness of the cloud manufacturing service platform and the processing speed of the system. In [29], this paper has proposed a novel P2P energy trading system for two separate optimization problems, one is an individual optimal charging algorithm designed for those consumers to obtain the best daily charging schedule, the other is a P2P energy trading mechanism to reduce the total daily energy cost. But they ignore the coupling between the two optimization problems. In [30], this paper build a trust mechanism based on blockchain technology, view the creation of digital assets as a process of evaluating behavior, design smart contracts to handle the evaluation behavior, and build a blockchain system based on the reputation values of alliance members. The system uses sidechain technology to transfer the created digital assets, which can increase the authenticity guarantee of the blockchain in other trading scenarios. Experimental results show that the system is characterized by low cost and memory space that is not easily expanded. However, this article does not evaluate the performance of the system.

Although some works have studied the system of electricity transaction, there exist new challenges to address. On the one hand, How to achieve better QoE and reliable services to meet the needs of consumers. On the other hand, how to guarantee the power supplier's reward and revenue maximization. More importantly, how to meet the above two requirements as far as possible, under the premise of the best service quality as far as possible to reduce the cost and expand the revenue, we will use the trusted reputation management system and the problem of revenue maximization two aspects to study.

We believe that reinforcement learning would be a good choice when the system needs to make decisions to achieve a balance between risk and reward in complex situations [31–33]. The selective federated reinforcement learning (SFRL) proposed in [34] can improve the accuracy of the automatic driving model very well. In this paper, we propose a blockchain-based smart grid system that can ensure efficient energy transactions by considering the situation of each node comprehensively, and can dynamically select the consensus mechanism of the blockchain to achieve consensus optimization. The efficient implementation of the system enables electricity consumers to obtain high quality and reliable energy services, while electricity suppliers are richly rewarded, thus motivating them to participate in the transactions again. For the real-time dynamics of the system, we design a MADDPG-based reinforcement learning algorithm to decide the electricity supplier that gets the power trading rights in each round, the generating power, and the consensus mechanism among the nodes. Multi-agent deep deterministic policy gradient (MADDPG) is a powerful reinforcement learning algorithm. In recent years, the leading contenders are deep Q-learning [35], "vanilla" policy gradient methods [36], and trust region natural policy gradient methods [37, 38]. However, Q-learning is not ideal in dealing with high-dimensional problems, because it is easy to be constrained by dimensional disasters and is poorly understood, vanilla policy gradient methods have poor data effiency and robustness.

The contributions of this paper are summarized as follows: (i) in order to solve the problems of mutual trust under the electricity transaction and realize the transparency of the system, we propose a reputation evaluation system based on blockchain technology, so that the power consumers can obtain reliable and better QoE services; (ii) in order to make the lowest cost of power suppliers and the biggest gains, we try to optimize transmitted power and charging power. We also choose the consensus algorithm and the state of the charge and discharge method to enable the power supplier to earn a larger profit and actively participate in the power supply and trading incentives. (iii) using reinforcement learning MADDPG effectiveness to solve the convergence of the experimental results.

The remainder of this paper is organized as follows. The related works are described in Section "Introduction". Section "System description and problem formulation" depicts the system model and problem formulation. Section "Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithms" presents the solution to the optimization problem. Section "Experiment result" presents the simulation results. The conclusion and future research issues are given in Section "Conclusion".

System description and problem formulation

System scenario

There is an electric power system with a set of electric consumers (ECs) $\mathcal{M} = \{1, 2, 3, \cdots, M\}$ and some electric suppliers (ESs) $\mathcal{N} = \{1, 2, 3, \cdots, N\}$. ESs can generate electricity through new energy sources such as wind energy, solar energy, and tidal energy, and can also generate electricity through conventional energy sources such as hydropower, oil, and nuclear energy. ECs can be different power-consuming users such as factories, residential life, and charging piles. At the beginning, ECs send the power order requests to ESs, ESs monitor the transaction requests, and N ESs compete for the transaction right of the orders at the same time. The system scenario is shown in Fig. 1. Assuming that the electronic request is an electronic order transaction, the electronic request is packaged into blocks and consensus is carried out in the blockchain. If the ES n is the first to obtain the power to produce blocks, it will obtain the right to trade electricity energy. The optimized strategy of transactions between EC and ES can be executed through smart contracts, thereby ensuring its correct, reliable and transparent execution.

Trust and reputation modeling

The paper [39] proposes a reputation management scheme based on multi-arm slot machine (MAB), which can effectively select vehicles with good reputation. In our system, there are two situations when two ESs interact. Communication trust generally refer to the transmission of data, including both cooperative and non-cooperative situations. The case of data trust generally refers to the data aggregation, including correct transmission and incorrect transmission. In Bayesian analysis, the beta distribution is usually used to represent the conjugate prior distribution of the binomial distribution parameters, where the beta distribution is simple and flexible, and can be used to simulate the trust distribution.

Beta function can be described by gamma function as follows,

$$P(x) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} x^{a-1}(1-x)^{b-1}, \ \forall\, 0 \le x \le 1, a \ge 0, b \ge 0, \tag{1}$$

where a represents the number of normal cooperation and b represents the number of data transmission error. For the prediction of ES's behavior, the probability distribution P of ES's reputation can be obtained by using beta distribution. When calculating trust and reputation values, we consider both communication trust and data trust. According to the beta distribution, the reputation of ES m to ES n in time slot t is expressed as [40]

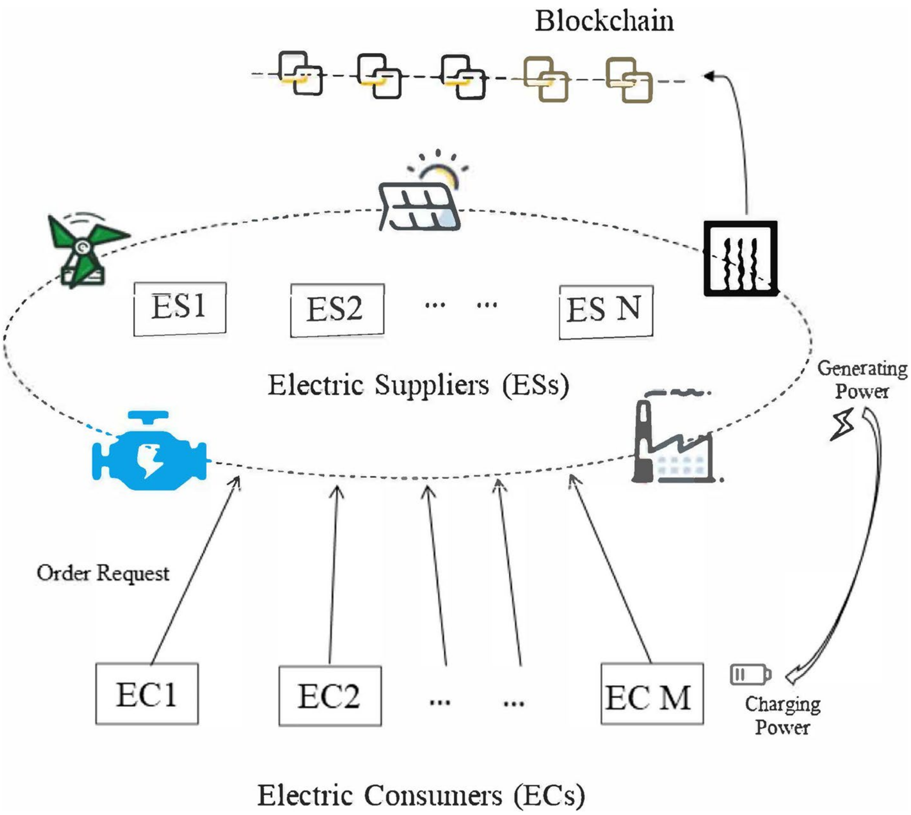

Fig. 1 The system scenario

$$T^{(t)}_{n,m} = \text{Beta}(a+1, b+1). \tag{2}$$

Therefore, we can obtain the final reputation value of ES n in time slot t,

$$T_n(t) = \sum_{m=1, m\neq n}^{M} T_{n,m}(t). \tag{3}$$

In our reputation scheme, we classify all ESs in the system into three categories, which are trusted nodes, uncertain nodes and untrustworthy nodes. For both trusted and uncertain nodes we will give the opportunity to participate in the transaction, while untrustworthy nodes need to stay in the network for observation. We will give all nodes an initial reputation value $T_{in} = 0.5$, which means that all nodes are initially uncertain nodes. The study of [41] points out that malicious nodes are a minority in P2P systems and suspicion of additional nodes is one of the important reasons for the degradation of the overall system performance. The formula given by [42]

$$TR_n(t) = T_{in} + \frac{T_n(t)}{M}, \tag{4}$$

is used to distinguish the type of Es n. When $\frac{T_n(t)}{M} < 0.5$, the reputation value is untrustworthy.If $\frac{T_n(t)}{M} = 0.5$, the reputation value is uncertain; if $\frac{T_n(t)}{M} > 0.5$, the reputation value is trustworthy. Trustworthy nodes can participate in the next round of transactions, while untrusted nodes need to stay in the network for observation. The more times a node is untrusted, the longer it will wait.

Since the deployment environment of each ES cannot be determined, some problems will inevitably arise when ESs are distributed in a harsh environment. Therefore, it makes more sense to provide a second chance for untrusted nodes. Untrusted nodes should stay in the network for observation rather than be immediately excluded from the transaction. When a node is considered untrusted, we mark it as untrusted and start the clock $T(t)$. During this period, untrusted nodes are not allowed to participate in power transactions. After this period, untrusted nodes will be restored to their initial credibility. $T(t)$ is not fixed. The more times nodes are untrusted, the longer nodes stay in the network for observation until nodes are blacklisted.

Blockchain system in energy trading

By deploying blockchain in the energy trading market, consensus algorithms and smart contracts can be used to make the entire trading process more reliable, credible, and transparent.

Consensus mechanism

ES uses different consensus algorithms to produce different block intervals and transaction throughputs. Denote $\beta(t,x)=\{0,1\}$ as the parameter of the consensus mechanism to show whether the consensus algorithm x is selected, where $x\in\{0,1,2\}$ represents the blockchain choosing the different consensus algorithms, PBFT, DPOS, and POS. $\beta(t,x)=1$ means consensus algorithm x is chosen. Otherwise, consensus algorithm x is not chosen in time slot t. These three commonly used consensus algorithms are described as follows,

(1) Practical Byzantine Fault Tolerance (PBFT): FBFT can tolerate not only node failures but also the existence of certain malicious nodes or Byzantine nodes. PBFT has requirements on the number of nodes in the system. Similar to the Byzantine Generals problem, PBFT requires that the number of nodes in the system N be no less than $3f+1$, where f is the number of "malicious nodes". The "malicious node" here can be a node that is deliberately malicious, a node that is attacked and controlled, or even a node that has lost its response. In short, as long as it is abnormal, it can be considered malicious. PBFT classifies each node in the system into two categories: primary node and replica nodes. They all use a state machine mechanism to record their actions. If the operation of each node is consistent, then their state machine will always remain consistent.

(2) Delegated Proof of Stake (DPoS): In the DPoS consensus algorithm, the normal operation of the blockchain depends on the delegates, and these delegates are completely equivalent. It is to vote through the proportion of stake, and more people have joined the power of the community. People will vote to select relatively reliable nodes for the maximization of their own interests, which is more secure and decentralized. DPOS uses a professionally run network server to ensure the security and performance of the blockchain network. It does not require computing power to solve mathematical problems, but the holder of the stake chooses who will say the producer.

(3) Proof of Stake (PoS): PoS is a consensus algorithm that distributes interest based on the amount and time of stake you hold. The core logic of the POS mechanism is that whoever holds the stake has control over the network. In the POS mechanism, there is still computing power mining, which requires computing power to solve a mathematical problem. However, the difficulty of mathematical problems is related to the "coin age" of the coin holder. The longer the coin holder has the coin, the simpler the problem and the greater the probability of mining the coin. The more stake it has, the greater the chance of meeting the Hash goal and obtaining the accounting right.

Generate energy trading blocks

ES n generates blocks according to the set block interval. Before this new block is added to the blockchain, it needs to be transferred to other ESs ($n^{'}\neq n$) for block verification. Let $I_b(t)$ and $T_b(t)$ denote the trading block size and block interval, respectively. We can separately obtain the block propagation time $T_p(t)$ and verification time $T_v(t)$.

$$T_p(t)=\max\left\{\frac{I_b(t)}{R_{n,n'}T_n(t)}\right\},\ T_v(t)=\max\left\{\frac{I_b(t)}{f^b_{n'}}\right\}, \tag{5}$$

where $R_{n,n'}$ is the transmission rate between ES n and $n^{'}$, and $f^b_{n'}$ is the clock speed of CPU consumed by verifying the block for ES $n^{'}$. We assume that ES has a first in first out (FIFO) data buffer to store the arrived but not yet verified blocks. Hence, the dynamics of the processing queue at the beginning of the $t+1$ time slot can be given by as follows,

$$F_{n'}(t+1)=\max\left\{F_{n'}(t)-f^b_{n'}\ ,\ 0\right\}. \tag{6}$$

Therefore, the total time cost in the consensus process can be given by

$$T_{total}(t)=T_b(t)+T_p(t)+T_v(t). \tag{7}$$

Then, the energy transaction throughput [43] can be expressed as

$$T_h(t)=\frac{\lfloor I_b(t)/\chi(t)\rfloor}{T_b(t)}, \tag{8}$$

where $\chi(t)$ is the average size of transactions.

Energy trading model

After the new block is added to the blockchain, the physical transaction of electricity between ES and EC can take place in the energy market. Let the transaction power be $P = \{P_{ge}^{n(t)}, P_{ch}^{m(t)}\}$, where $P_{ge}^{n(t)}$ is the generating power of ES n in t time slot, and $P_{ch}^{m(t)}$ is charging power of EC m. Let $x_n(t) \in \{0, 1\}$ be the power supply status of the ES n. Specifically, $x_n(t) = 1$ is the state of selling electricity, and $x_n(t) = 0$ means stop the power supply.

In order to optimize the revenue of suppliers ESs and incentivize each ES to participate in electricity supply in blockchain-enabled smart grids, the benefit of ES n can be given by

$$R_n(t) = T_h(t)\rho_m(t)P_{ch}^{m(t)} - C_1(P_{ge}^{n(t)}) - C_2(P_{ge}^{n(t)}), \tag{9}$$

where $C_1(P_{ge}^{n(t)})$ is the operating cost of ES n to generate power $P_{ge}^{n(t)}$ in time slot t, and $C_2(P_{ge}^{n(t)})$ is the basic maintenance cost of ES n to generate power $P_{ge}^{n(t)}$ in time slot t. $\rho_m(t)$ is the unit payment by the EC m for the obtained charging power $P_{ch}^{m(t)}$ from ES n with different reputations.

Problem formulation

In order to allow ECs to obtain charging services from highly reliable and high-quality ESs, while ensuring the benefits of ESs, so as to realize a virtuous circle of energy trading market. We can get the following utility function in time slot t,

$$U(t) = E\left[\sum_{t=1}^{T-1} \omega_1\omega_2 \sum_{n=1}^{N} x_n^t R_n(t) + (1-\omega_1)T_h(t)\right], \tag{10}$$

where $\omega_1 (0 < \omega_1 < 1)$ is a weight factor to combine the benefit $R_n(t)$ and throughput of the blockchain $T_h(t)$, and ω_2 is a mapping factor that ensures that the two functions is at the same level.

Let $A = \{P_{ge}^{n(t)}, P_{ch}^{m(t)}, \beta(t, x), x_n(t)\}$, eventually, we can formulate the following energy trading and consensus optimization problem to maximize the benefit of electric consumers,

$$\begin{aligned}
&\max_{A} \ U(t) \\
&\quad \text{s.t.}\ (C_1): P_{ge}^{min} \le P_{ge}^{n(t)} \le P_{ge}^{max}, \\
&\qquad (C_2): \beta(t, x) \in \{0, 1\}, x \in \{0, 1, 2\}, \\
&\qquad (C_3): \beta(0) + \beta(1) + \beta(2) = 1, \\
&\qquad (C_4): T_{total}(t) \le T_b(t), \\
&\qquad (C_5): 0 \le P_{ch}^{m(t)} \le P_{ch}^{max},
\end{aligned} \tag{11}$$

where P_{ge}^{min}, P_{ge}^{max}, and P_{ch}^{max}, are the minimum generating power, the maximum generating power, and maximum charging power, respectively. z^{max} is the maximum delay requirement for block generation, and we can set the block interval as $T_b(t) = f(\beta(t, x), z^{max})$ in slot t, which is caused by the selection of different consensus algorithms x. In (11), (C_1) and (C_5) are the generating power and charging power constraints, respectively. (C_2) and (C_3) restrict the consensus algorithm selections. (C_4) is the constraint of the total time delay of generating block.

Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithms

In the system, there are MECs [44, 45] and NESs in each round, and the ECs initiate the power trade request and the ESs compete for the power trade right. In order to ensure real-time and reliable transactions, the system selects the status of ESs, generating power(P_{ge}) of ESs, and consensus mechanism of the blockchain of the competing ESs in each round. This selection problem can be modeled as MDP.

Considering that the system needs to carry a huge volume of transactions, we use a DRL algorithm known as MADDPG as the solution. MADDPG is based on deep deterministic policy gradient(DDPG). MADDPG improves the actor-critic framework, which adopts the rule of centralized training and decentralized execution. It provides a general and novel idea for solving multi-agent problems. Firstly, similar to RL, agents interact with the environment according to the principles of MDP and receive rewards. The purpose is to continuously accumulate experience to make better decisions adapting to the environment. Specifically, during slot t, the agent will update its state-value function according to $(S_n(t), A_n(t), r_n(t), S_n(t+1))$ as follows:

$$Q(S_n(t), A_n(t)) \leftarrow Q(S_n(t), A_n(t)) + \alpha\sigma(t). \tag{12}$$

Then we expand $\alpha\sigma(t)$ as follows:

$$\alpha[R(t+1) + \gamma Q(S_n(t+1), A_n(t+1)) - Q(S_n(t), A_n(t))]. \tag{13}$$

Where $\sigma(t)$ is the TD error, which should be 0 at the best Q value, α is the learning rate, and γ is the discount factor that narrows with t increases. After the expansion of the TD error, $R(t+1) + \gamma Q(S_n(t+1), A_n(t+1))$ is the TD Target, which minus the current $Q(S_n(t), A_n(t))$ to get TD error, which can be understood as the updated value of Q. The Target value in the formula can be expanded as follows:

$$Target = U_t = R(t+1) + \gamma R(t+2) + \gamma^2 R(t+3) + \ldots \tag{14}$$

Which means the sum of expected future rewards. *Target* is not known in slot t. But we calculate it by:

$$Q_\pi(S_n(t), A_n(t)) = \mathbb{E}(U_t | S_n(t), A_n(t)). \tag{15}$$

This formula eliminate the *State* and *Action* after $t+1$. And Q_π is called the action-value function. Furthermore, the maximum value of Q_π can be obtained by:

$$Q^*(S_n(t), A_n(t)) = \max_\pi Q_\pi(S_n(t), A_n(t)). \tag{16}$$

Consider a Markov decision process, defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}', \mathcal{G})$ representing the dynamics of the system.

State $\boldsymbol{\mathcal{S}}$ **:** Space of states of the system, which are the input of the actor network. The state of system in time slot t is denoted by $S(t)$,$S(t) \in \mathcal{S}$.Define $S(t)$ as follows:

$$S(t) = [T(t), \Phi_s(t), F(t)], \tag{17}$$

where $T(t)$ and $\Phi_s(t)$ are the reputations and the stakes of blockchain nodes in time slot t, respectively. Denote the sets of reputation and the sets of the stake by $T(t) = \{T_1(t), T_2(t), ..., T_N(t)\}$ and $\Phi_s(t) = \{\Phi_1(t), \Phi_2(t), ..., \Phi_N(t)\}$, respectively. $F(t)$ is the computing resources of edge servers,which generate blocks and verify transactions. There are as many edge servers in the system as there are ES. Denote the sets of computing resource of edge servers by $F(t) = \{F_1(t), F_2(t), ..., F_N(t)\}$. The computing resource of edge server n in time slot $t+1$ can be given by (6).

Action $\boldsymbol{\mathcal{A}}$ **:** Space of actions, $a(t) \in \mathcal{A}$. Let $a(t)$ denote the action selected by the actor network in time slot t. Define $a(t)$ as follows:

$$a(t) = [x(t), P_{ge}(t), \beta(t)], \tag{18}$$

where $x(t) = \{x_1(t), x_2(t), ..., x_N(t)\}$, $P_{ge}(t) = \{P_{ge}^{1(t)}, P_{ge}^{2(t)}, ..., P_{ge}^{N(t)}\}$, $\beta(t) = \{\beta(t,0), \beta(t,1), \beta(t,2)\}$.

$\boldsymbol{\mathcal{P}'}$ **:** A state transition probability matrix. $P'(s(t+1)|s(t), a(t))$ defines the probability that the state $s(t)$ transforms to $s(t+1)$ under action $a(t)$.

$\boldsymbol{\mathcal{G}}$ **:** The total discounted return from t to $t+j$ can be expressed as:

$$G(s(t)) = r(t) + \gamma r(t+1) + ... + \gamma^j r(t+j) = \sum_{i=0}^{j} \gamma^i r(t+i), \tag{19}$$

where $\gamma \in (0, 1]$ is a discount factor that encodes the importance of future rewards, and $r(t)$ denotes the rewards available in the current state $s(t)$:

$$r(t) = \begin{cases} U(t) & \text{if } C_1, C_2, C_3, C_4, C_5 \textit{ are satisfied} \\ 0 & \textit{otherwise} \end{cases} \tag{20}$$

***MADDPG Method* :**

The algorithm is as shown in Algorithm 1. In line 4, we use $\epsilon - greedy$ to select a random action. in line 5-line 6, each Con_n executes the action and receive a reward and then load them to the replay buffer. In line7-line 11, each Con_n updates its actor and critic and target network. Specifically, We let $Con_n, n \in N$ act as agents. And we use $\theta = [\theta_1, \theta_2, ..., \theta_n]$ represent the policy parameters of agents. Then we use $\pi = [\pi_1, \pi_2, ..., \pi_n]$ represent the policies of agents, each agent updates its policy parameters to obtain the optimal target policy $\pi^*_{\theta_n} = \arg\max_{\theta_n} J(\theta_n)$. For the deterministic policy gradient algorithm on continuous action space, the actor will output deterministic greedy actions according to the state, which may lead to some actions never being chosen, so the random behavior policy must be used to ensure adequate exploration when selecting actions. So we should use the random policy to get actions as much as possible. Here we use $\epsilon - greedy$ to explore actions:

$$\pi_n(a_n|o_n) = \begin{cases} \arg\max_{\theta_n} J(\theta_n), & \textit{with probability } 1-\epsilon \\ rand(a_n), & \textit{with probability } \epsilon \end{cases} \tag{21}$$

With the training progresses, ϵ is gradually reduced to 0. So the final result is still a deterministic policy. During the train, actor updates the policy by calculating the gradient of $J(\pi_n)$. This deterministic policy gradient formula is as follows:

$$\nabla_{\theta_n} J(\pi_n) = \mathbb{E}_{\mathbf{o},a\sim D}\left[\nabla_{\theta_n}\pi_n(a_n|o_n)\nabla_{\theta_n} Q_n^\pi(\mathbf{o}, a_1, a_2, ..., a_n)\right]. \tag{22}$$

Where $a_n = \pi_n(o_n)$. $\mathbf{o} = \{o_1, o_2, ..., o_n\}$ is the local observation for the agent. And $Q_n^\pi(\mathbf{o}, a_1, a_2, ..., a_n)$ is the centralized action-value function of the agent. Each agent learns its own Q_n^π independently and obtains rewards, so agents can complete the cooperative task in this model. D is an experience replay buffer which is composed of $(\mathbf{o}, \mathbf{o}', a, r)$. In addition, the centralized critic updates the action-value function Q_n^π according to the following minimization loss function:

$$L(\theta_n) = \mathbb{E}_{\mathbf{o},\mathbf{o}',a,r}\left[(Q_n^\pi(\mathbf{o}, a_1, a_2, ..., a_n) - y_n)^2\right]. \tag{23}$$

Where, $y_n = r_n + \gamma \overline{Q_n^\pi}(\mathbf{o}', a_1', a_2', ..., a_n')|_{a_n' = \pi_n'(o_n)}$ is the TD Target. $\overline{Q_n^\pi}$ is the target network. And $\pi_n' = \left[\pi_1', \pi_2', ..., \pi_n'\right]$ is the parameter that the target policy has lagged update property. At the end of each train, the agent will get the learned policy parameters and updates its own actor and critic network parameters by:

$$\theta_n' \leftarrow \zeta\theta_n + (1-\zeta)\theta_n'. \tag{24}$$

Where ζ is the update step.

The process of MADDPG-based ESs state, P_{ge} and consensus mechanism selection algorithm is shown in Algorithm 1.

```
Input: Learning rate $r$, exploration probability $\epsilon$, discount
    factor $\gamma$, update step $\zeta$, replay buffer size $\mathcal{D}$
Output: Max reward $r^*$, optimal policy $\pi^*_{\theta_n}$
    Initialize $Q(s,a)$, $State_n$, $Action_n$, $Reward_n$
for $episode = 1 \rightarrow \infty$ do
    for $n = 1 \rightarrow N$ do
        Select an action $a_n(t)$ by $\epsilon - greedy$ strategy and
            execute $a_n(t)$
        Observe the system reward $r_n(t)$ and the new state
            $s'$
        Store $s, a, r, s'$ into replay buffer
        Sample a batch of record from replay buffer
        Update behavior critic by minimizing the loss via
            formula (23)
        Update actor by using the sampled policy gradient
            via formula (22)
    end for
    Update the target network parameters for each agent via
        formula (24)
end for
```

Algorithm 1 The procedure of MADDPG in system

Experiment result

In this section, we exhibit the performance of the energy trading and consensus optimization.

Simulation parameters

We simulate the performance of the processed scheme based on Prtorch 1.0.2 with Python 3.9 as the software environment. The settings of the simulation parameters are shown below. we consider a energy trading system consisting of 20 ECs and 5 ESs. The function of the block interval is modeled as $\log(1 + \beta(t)z^{max})$. The minimum generating power, maximum generating power, and maximum charing power are respectively $P_{ge}^{min} = 0.2$ W, $P_{ge}^{max} = 2$ W, and $P_{ch}^{max} = 1$ W. The maximum time of the consensus algorithm is $z^{max} = 2$ s. Meanwhile, we show the three benchmark schemes to verify the proposed scheme. The first is the fixed consensus algorithm scheme, where one consensus algorithm is selected, referred as FCAS. The second is that the scheme dose not allocation the generating power of ES, called FPAS. The final is the single objective optimization, where the optimization problem only considers the benefits of ES, referred as SOAC.

Numerical results

In Fig. 2, we show the convergence of the resource allocation scheme based on the MADDPG algorithm. Observing the figure, we can find that the algorithm has a fast the convergence rate. Figure 3 shows the impact of the total reward of the system on the maximum

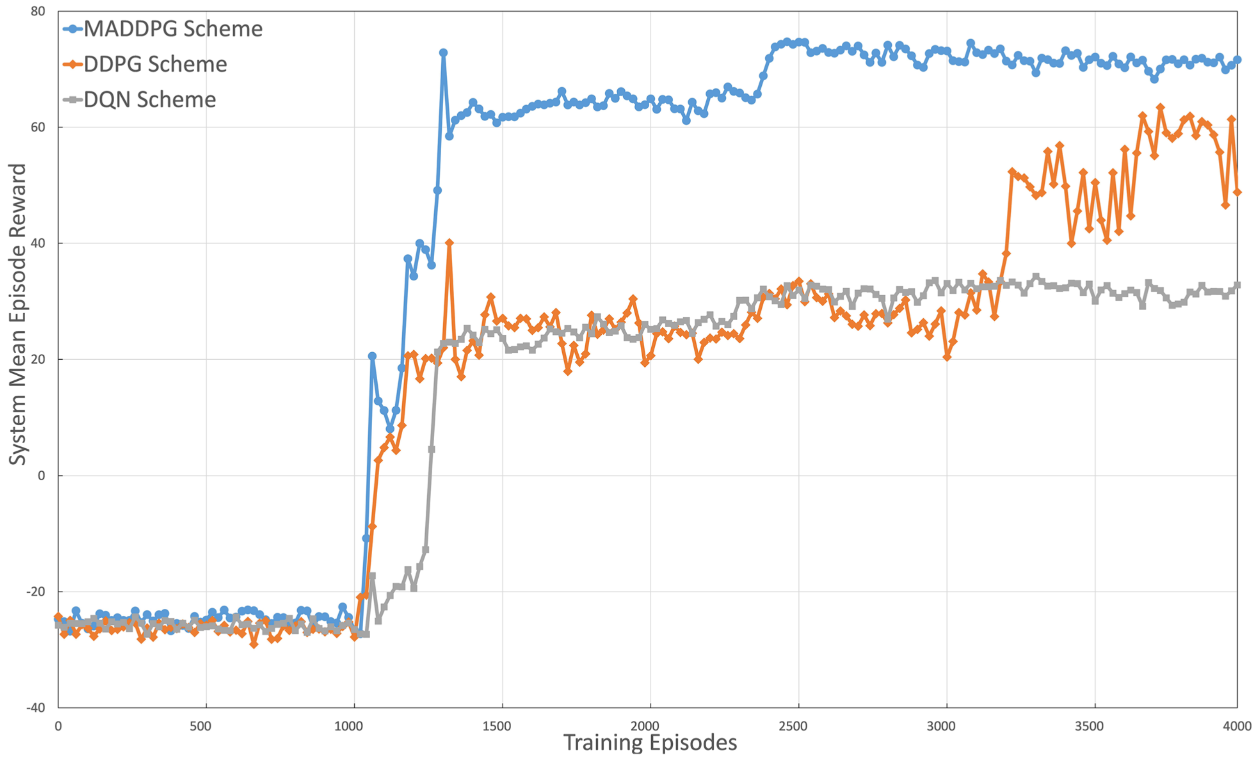

Fig. 2 Convergence of Algorithm 1

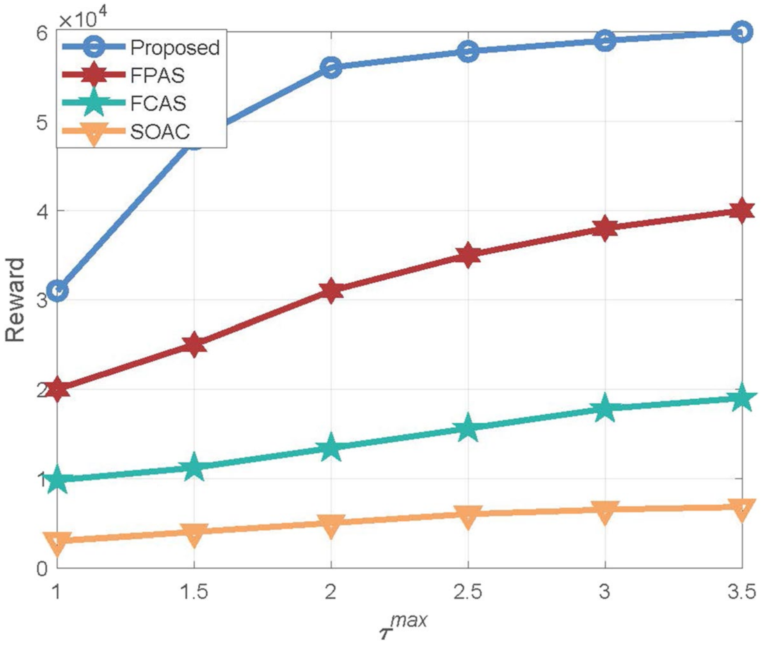

Fig. 3 Reward v.s. τ^{max}

delay requirement for block generation τ^{max}. It can be seen that the reward increase with the increase in the maximum delay requirement. Meanwhile, we find that the proposed scheme has the best performance, while the SOCA scheme has the worst performance. This is because the single-objective optimization does not consider dynamic edge computing node resource changes and competition.

In Fig. 4, we show the effect between reward and P_{ch}^{max}. Looking at the trend of the graph, we can find that the increase of P_{ch}^{max} has a growing trend in the influence of reward. Obviously, the proposed energy trading scheme

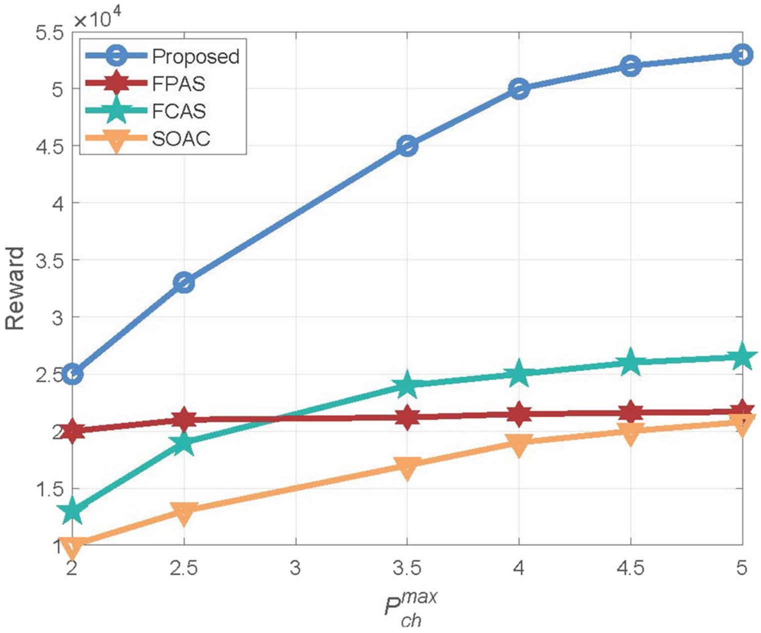

Fig. 4 Reward v.s. P_{ch}^{max}

performs the best. At the same time, we find that the growth of P_{ch}^{max} has little effect on FPAS, because FPAS itself does not allocate power, so P_{ch}^{max} does not affect the performance of the FPAS scheme . The impact of P_{ch}^{max} on the other two schemes, namely FCAS and SOCA, is relatively small. Figure 5 shows the effect of the number of ESs on reward. From the figure, we can see that as the number of ES increases, the reward shows an increasing trend.

In Fig. 6, we show the effect between average selection ratio and reputation value $T(t)$. Obviously, we can find that as the ES reputation value increases, the ES is more likely to obtain power trading rights, and ESs with low reputation value will not be completely deprived of the

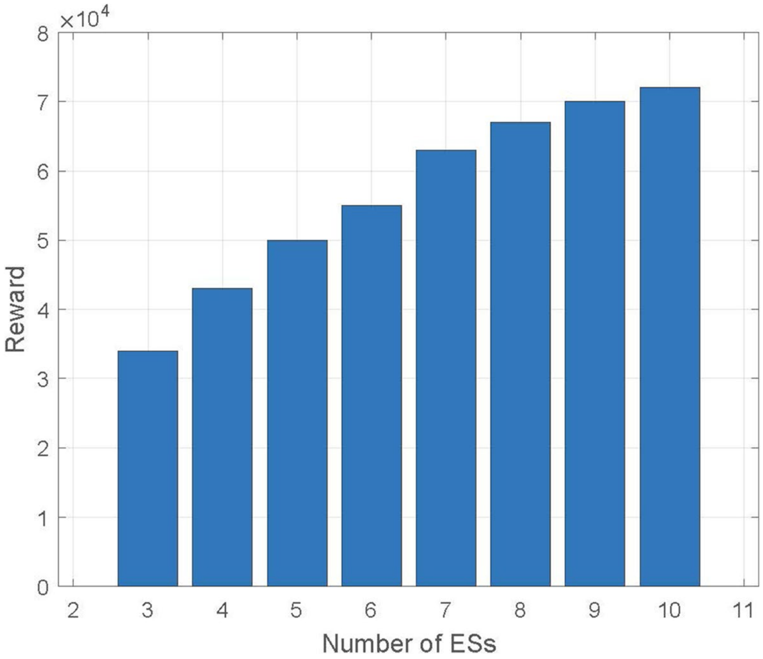

Fig. 5 Reward v.s. Number of ESs *N*

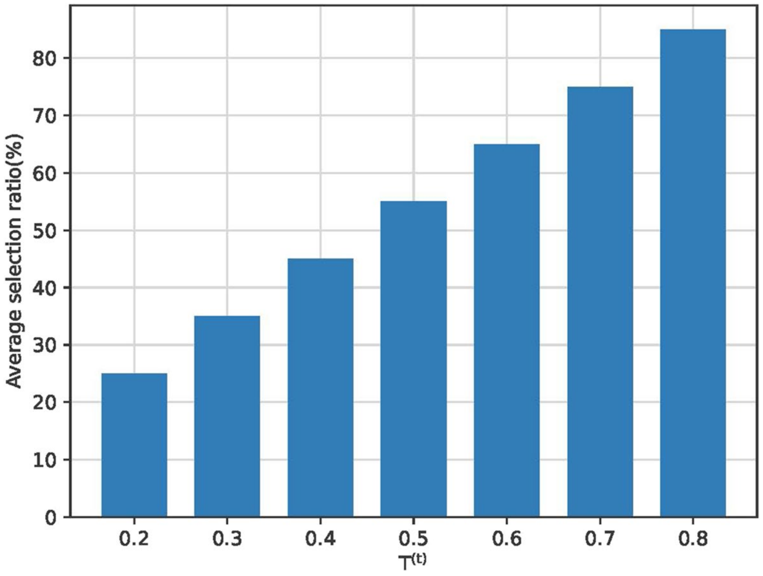

Fig. 6 Average selection ratio v.s. Reputation value *T*(*t*)

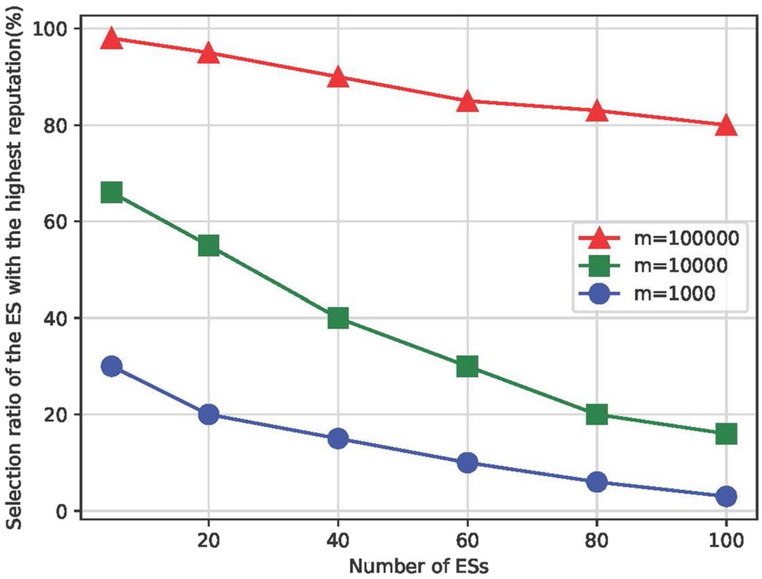

Fig. 7 Selection ratio of the ES with the different trading rounds *m* v.s. Number of ESs *N*

power to obtain power trading rights. Figure 7 shows the effect between selection ratio of the ES with the different trading rounds *m* and the number of ESs *N*. We can find that, for a given number of transactions, as the number of ESs increases from 5 to 100, the proportion of ESs with high reputation value to obtain power trading rights gradually decreases. This is because a high reputation is not the only requirement for gaining power trading rights. The system will give ES with low reputation value the opportunity to participate in the transaction. As the number of transaction rounds m increases, the proportion of high-reputation ESs obtaining power trading rights will also increase, because at this time the reputation value of each ES is in a relatively stable state, and the system has accumulated enough experience to make the best choice.

Conclusion

Although distributed energy has become a hot research topic, there are still many problems in distributed energy trading system, such as user privacy protection and mutual trust in trading, how to ensure the high quality and reliability of energy services, how to encourage energy suppliers to participate in transactions. To solve these problems, in this paper, we propose a blockchain smart grid system to optimize efficient energy transactions and blockchain consensus using a reinforcement learning MADDPG algorithm for power supplier selection. Through the construction of a reputation evaluation system, electricity consumers can obtain reliable and high-quality power services. In addition, the generation and charging power are optimized in this paper. By choosing the consensus algorithm and charging and discharging states, the power supplier's revenue is maximized, thus incentivizing the power supplier to participate in the supply trading network and ensuring the long-term stability of the power resource market. Finally, we analyze the simulation results in detail and compare them with existing algorithms. The feasibility of the proposed algorithm can be demonstrated by the validity and convergence of the results.

Authors' contributions

Ruohan Wang and Yunlong Chen conducted the system model and algorithm design, Hongwei has written the main manuscript text; Entang Li and Jing Li performed the simulation; and Xueyao Zhang conducted the performance analysis.

Funding

Marketing Program of State Grid Shandong Electric PowerCompany (640633210015).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 31 March 2023 Accepted: 5 August 2023
Published online: 19 August 2023

References

1. Zeng M, Zhang X, Wang L (2016) Energy supply side reform promoting based on energy internet thinking. Electric Power Constr 37(4):10–15
2. Qazi A, Hussain F, Rahim NA, Hardaker G, Alghazzawi D, Shaban K, Haruna K (2019) Towards sustainable energy: a systematic review of renewable energy sources, technologies, and public opinions. IEEE Access 7:63837–63851
3. Qili H (2019) Development road of green energy. Distrib Energy Resour 4(2):1–7
4. Olabi A, Abdelkareem MA (2022) Renewable energy and climate change. Renew Sust Energ Rev 158:112111
5. Gielen D, Boshell F, Saygin D, Bazilian MD, Wagner N, Gorini R (2019) The role of renewable energy in the global energy transformation. Energy Strateg Rev 24:38–50
6. Soeiro S, Dias MF (2020) Renewable energy community and the european energy market: main motivations. Heliyon 6(7):e04511
7. Hu J, Moorthy SK, Harindranath A, Zhang Z, Zhao Z, Mastronarde N, Bentley ES, Pudlewski S, Guan Z (2023) A mobility- resilient spectrum sharing framework for operating wireless UAVs in the 6 GHz band. IEEE/ACM Transactions on Networking. https://doi.org/10.1109/TNET.2023.3274354
8. Koch C, Hirth L (2019) Short-term electricity trading for system balancing: An empirical analysis of the role of intraday trading in balancing germany's electricity system. Renew Sust Energ Rev 113:109275
9. Chen G, Li M, Xu T, Liu M (2017) Study on technical bottleneck of new energy development. Proc CSEE 37(1):20–26
10. Xutao G, Jiecong C, Gaoyan H, Na X, Hongkun L (2019) Technologies and development status for distributed energy resources. Distrib Energy Resour 4(1):52–59
11. Li C, Li Z, Zhu H, Tian Z, Feng W (2020) Study on operation strategy and load forecasting for distributed energy system based on chinese supply-side power grid reform. Energy Built Environ 3(1):113–127
12. Georgilakis PS (2020) Review of computational intelligence methods for local energy markets at the power distribution level to facilitate the integration of distributed energy resources: State-of-the-art and future research. Energies 13(1):186
13. Bayram IS, Shakir MZ, Abdallah M, Qaraqe K (2014) A survey on energy trading in smart grid. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE, pp 258–262
14. Liu Y, Wu L, Li J (2019) Peer-to-peer (p2p) electricity trading in distribution systems of the future. Electr J 32(4):2–6
15. Abdella J, Shuaib K (2018) Peer to peer distributed energy trading in smart grids: A survey. Energies 11(6):1560
16. Li G, Li Q, Song W, Wang L (2021) Incentivizing distributed energy trading among prosumers: A general nash bargaining approach. Int J Electr Power Energy Syst 131:107100
17. Wang Q, Su M, Li R, Ponce P (2019) The effects of energy prices, urbanization and economic growth on energy consumption per capita in 186 countries. J Clean Prod 225:1017–1032
18. Fu Y, Li C, Yu FR, Luan TH, Zhao P, Liu S (2023) A survey of blockchain and intelligent networking for the metaverse. IEEE Internet Things J 10(4):3587–3610
19. Xiong W, Binyou Y, Zhang R et al (2020) Research ondistributedenergytradingmodelbasedonconsortiumchain. SmartPower 48(10):24–29
20. Xu J, Wen M, Zhang K, Chen X (2019) Bidding transaction platform for distributed electrical energy based on blockchain. Xi'an, Smart Power 47(10):56–62
21. Shen X, Pei Q-Q, Liu X-F (2016) Survey of block chain. Chin J Netw Inf Secur 2(11):11–20
22. Chong PHJ, Seet B-C, Chai M, Rehman SU (2018) Smart Grid and Innovative Frontiers in Telecommunications: Third International Conference, SmartGIFT 2018, Auckland, New Zealand, April 23-24, 2018, Proceedings, vol. 245. Springer
23. Che Z, Wang Y, Zhao J, Qiang Y, Ma Y, Liu J (2019) A distributed energy trading authentication mechanism based on a consortium blockchain. Energies 12(15):2878
24. Liu C, Chai KK, Lau ET, Chen Y (2018) Blockchain based energy trading model for electric vehicle charging schemes. In: International Conference on Smart Grid Inspired Future Technologies. Springer, pp. 64–72
25. Feng J, Zhang W, Pei Q, Wu J, Lin X (2022) Heterogeneous computation and resource allocation for wireless powered federated edge learning systems. IEEE Trans Commun 70(5):3220–3233
26. Feng J, Liu L, Pei Q, Li K (2022) Min-max cost optimization for efficient hierarchical federated learning in wireless edge networks. IEEE Trans Parallel Distrib Syst 33(11):2687–2700
27. Zhang N, Wang Y, Kang C, Cheng J, He D (2016) Blockchain technique in the energy internet: preliminary research framework and typical applications. Proceedings of the CSEE 36(15):4011–4022
28. Wang Q, Liu C, Zhou B (2019) Trusted transaction method of manufacturing services based on blockchain. Comput Integr Manuf Syst 25(12):3247–3257
29. Alvaro-Hermana R, Fraile-Ardanuy J, Zufiria PJ, Knapen L, Janssens D (2016) Peer to peer energy trading with electric vehicles. IEEE Intell Transp Syst Mag 8(3):33–44
30. Wang X, Wang J, Zhang Y et al (2018) Blockchain system for creating digital assets based on reputation value. Netinfo Secur 5:59–65
31. Du J, Cheng W, Lu G, Cao H, Chu X, Zhang Z, Wang J (2022) Resource pricing and allocation in MEC enabled blockchain systems: An a3c deep reinforcement learning approach. IEEE Trans Netw Sci Eng 9(1):33–44
32. Shi T, Cai Z, Li J, Gao H, Qiu T, Qu W (2022) An efficient processing scheme for concurrent applications in the iot edge. IEEE Trans Mob Comput
33. Du J, Yu FR, Lu G, Wang J, Jiang J, Chu X (2020) MEC-assisted immersive VR video streaming over terahertz wireless networks: A deep reinforcement learning approach. IEEE Internet Things J 7(10):9517–9529
34. Fu Y, Li C, Yu FR, Luan TH, Zhang Y (2023) A selective federated reinforcement learning strategy for autonomous driving. IEEE Trans Intell Transp Syst 24(2):1655–1668
35. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
36. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: International conference on machine learning. PMLR, pp 1928–1937
37. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: International conference on machine learning. PMLR, pp 1889–1897
38. Wu M, Yu FR, Liu PX, He Y (2022) A hybrid driving decision-making system integrating markov logic networks and connectionist AI. IEEE Trans Intell Transp Syst
39. Xiao H, Cai L, Feng J, Pei Q, Shi W (2023) Resource optimization of mab-based reputation management for data trading in vehicular edge computing. IEEE Trans Wirel Commun 1
40. Fang W, Zhang C, Shi Z, Zhao Q, Shan L (2016) Btres: Beta-based trust and reputation evaluation system for wireless sensor networks. J Netw Comput Appl 59:88–94. https://www.sciencedirect.com/science/article/pii/S108480451500140X
41. Friedman* EJ, Resnick P, (2001) The social cost of cheap pseudonyms. J Econ Manag Strateg 10(2):173–199
42. Fang W, Zhang C, Shi Z, Zhao Q, Shan L (2015) Btres: Beta-based trust and reputation evaluation system for wireless sensor networks. J Netw Comput Appl 59:88–94
43. Liu M, Yu FR, Teng Y, Leung VC, Song M (2019) Performance optimization for blockchain-enabled industrial internet of things (IIOT) systems: A deep reinforcement learning approach. IEEE Trans Ind Inform 15(6):3559–3570
44. Wu M, Yu FR, Liu PX (2022) Intelligence networking for autonomous driving in beyond 5g networks with multi-access edge computing. IEEE Trans Veh Technol 71(6):5853–5866
45. Shi T, Cai Z, Li J, Gao H, Chen J, Yang M (2022) Services management and distributed multihop requests routing in mobile edge networks. IEEE/ACM Trans Networking 31(2):497–510

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.