Analysis of Customer Behavior Characteristics and Optimization of Online Advertising Based on Deep Reinforcement Learning

Zhenyan Shang, Bi Ge
Chongqing College of International Business and Economics, Chongqing 401520, China

***Abstract*—With the shift from traditional media to online advertising, real-time strategies have become crucial, evolving to meet contemporary demands. Advertisers strive to succeed in online advertising evaluations by demand-side platforms to secure display opportunities. Discrepancies in information evaluation can impact click-through rates, emphasizing the need for precise prediction models in asymmetric contexts. Time dynamics significantly influence online ad click-through rates, with rest hours outperforming working hours. This study introduces the ARMA model to refine click predictions by preprocessing hits and employing a single XGBoost model. Furthermore, a reinforcement learning model is developed to explore online advertising strategies amidst information imbalances. Data is segmented into training (70%), validation (15%), and test sets (15%), with model parameters optimized using the DQN algorithm over 48 hours. Validation and testing on separate datasets comprising 15,000 entries each yield model accuracies of 0.85 and recall rates of 0.82. The incorporation of regret minimization algorithms enhances reward functions in deep reinforcement learning. Leveraging Tencent data, a comparative analysis evaluates advertisers' click rates as overrated, underrated, or accurately predicted by DSPs. Findings indicate that smart customer behavior characteristics outperform DQN, converging swiftly to optimal solutions under complete information. Smart characteristics exhibit stability and flexibility, with human-machine collaboration circumventing the drawbacks of random exploration. Transfer Learning amalgamates experimentation with real-world insights, bolstering algorithm adaptability for intelligent decision-making tools in enterprises.**

Keywords*—Real-time online advertising; ARMA-XGBoost model; information asymmetry; deep reinforcement learning decision-making behavior; Transfer Learning

I. Introduction

The publishing platforms and advertising types of online advertising showed a positive trend of development. Online advertising with the Internet as the media, provides advertisers with a high-yield and low-cost way of delivery [1, 2]. Compared with expensive display signs and paper advertisements, online advertising, through big data matching technology based on the characteristics of the audience, can deliver the most accurate revenue for advertisers. At the same time, through the identification of user intention, online advertising can provide users with more interesting advertisements, achieving a win-win situation for users and advertisers [3, 4]. At present, online advertising is generally divided into three categories: sponsored search advertising, general display advertising and real-time online advertising. In general, sponsored search ads are ads displayed in search results after users query their keywords on search engines such as browsers [5]. The general display advertisement will pop up when the user brow the website information, or when the user uses the mobile App, in the open screen animation of the software or the rotation map at the top of the home page [6]. The principle of RTB advertising is that advertisers design online advertising strategies, and through the demand side platform, online advertising on the web or the mobile App, to realize the advertisers to choose the corresponding advertising audience [7, 8].

RTB advertising has experienced explosive growth since its birth. In 2011 internationally, 88% of North American advertisers switched to RTB ads when they bought online ads. The RTB market is expected to grow to $9 billion in 2023, or 40% of the total advertising budget. In China, the RTB market first started with the TANX system launched by Taobao in 2011. By 2013, the number of RTB AD requests in China had reached five billion, and the RTB investment budget for advertisers had increased by 300% to $83 million [9]. RTB advertising is widely used in the era of mobile Internet. Compared with sponsored search advertising and general display advertising in the PC era, RTB advertising has changed the pattern of online advertising to a large extent. Every time in online advertising, you can accurately target the potential user audience, and select the most appropriate ads from the advertising library [10]. RTB iterates on data that is more relevant to user habits than focusing only on context keywords. Advertisers urgently need to make a more accurate prediction and evaluation of the display effect of advertising and design an appropriate online advertising model according to the estimated display effect. The display effect of advertisements can be evaluated by the conversion rate. The better the estimated display effect, the higher the cost of online advertising [11, 12]. Whether we can accurately predict the click rate is the key to the accuracy and effectiveness of the online advertising model. At the same time, in the actual online advertising process, multiple advertisers may participate in the online advertising auction. DSP ranks shot ads based on the effective click cost, and the highest ranked advertisers pay according to the broad second online advertising mechanism [13, 14].

In real, online advertising auctions, the price of online advertising needs to meet the cost constraints of the budget. How to accurately estimate the AD click-through rate, combined with the estimated click-through rate, participate in online advertising within the limited budget, win the DSP online advertising

evaluation has great research significance. In the background, the prediction and bidding model of real-time online advertising advertisements under asymmetric information [15]. After users see the ads they are interested in, they may click, download, register, buy and a series of behaviors, that is, the advertising effect has been transformed. Whether users respond to ads is one of the most concerned issues on the demand side. Users' download, registration and purchase behavior may cause longer delays, so scholars pay more attention to click-through prediction models, both in industrial applications and in academia [16]. In general, the click-through rate prediction model not only predicts the probability of users clicking after seeing an AD, but also describes how much the user is interested in the AD. Early AD click-through rate prediction models can be roughly divided into two categories: feature-based click-through rate prediction models and maximum-likelihood-based click-through rate prediction models. In feature-based methods, the prediction models are constructed based on page display features [17]. These features may include the text of the AD, the picture content, the location on the page, etc. Feature-based methods usually utilize logistic regression models. It describes the click and non-click behavior of ads as a dichotomy problem and is solved with a logistic regression model. Through the experiment, the model has a good prediction effect on the repeatedly displayed advertisements, and the accuracy of the prediction has increased steadily with the increase of the number of repeated displays [18, 19]. However, LR model is only effective for first-order features, and the ability to learn sparse features is poor. It relies on manual preprocessing of combined features, which should meet the shortage of complex data in practical applications.

II. Information Asymmetry in Real-Time Online Advertising

A. Information Asymmetry Analysis in Real-Time Online Advertising

Maximum-likelihood-based methods attempt to smooth the response estimates using the advertised exposure and hits, such as the Gamma-Poisson model. As shown in Eq. (1) and Eq. (2), these methods are all based only on simple linear models and cannot capture the associations between the data. In this case, a hybrid method is proposed by combining the hierarchical information of advertisements and the display information through matrix decomposition using the display features of advertisements.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \xi_t \tag{1}$$

$$z^p - \phi_1 z^{p-1} - \phi_2 z^{p-2} - \cdots - \phi_p = 0 \tag{2}$$

This method uses MF to learn a set of latent features from the data while correcting the prediction results with a feature-based approach. However, the MF used in this method is limited to binary relationships and does not satisfy higher-order relationships. As shown in Eq. (3) and Eq. (4), in the factorization machine model, the implicit vector of the first-order combined features is calculated to obtain the weight of the second-order combined features.

$$\mathrm{Cov}(\varepsilon_t, \varepsilon_s) = \sigma^2 \delta_{t-s} = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s, \end{cases} \tag{3}$$

$$\hat{\rho}_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x}_N)(x_{t+k} - \bar{x}_N)}{\Sigma_{t=1}^{N} (x_t - \bar{x}_N)^2} \tag{4}$$

The combination problem under sparse features is further solved. The concept of "field" is proposed on the FM model, classifying the features of the same properties into the same field and learning the hidden vector for each field. When learning different combinations of features, the internal product, as shown in Eq. (5) and Eq. (6), as the weight of the combined features. The experiment proves that FMM model can more accurately estimate the click rate of online ads compared with FM model.

$$LB = n(n+2) \sum_{k=1}^{m} \left(\frac{\hat{\rho}_k^2}{n-k} \right) \tag{5}$$

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{6}$$

Although FM and FMM models can solve the combined characteristics well, their theoretical essence is still the second-order combined characteristics stage, so they are not much used in the industry. As shown in Eq. (7) and Eq. (8), when the application of the click-through rate prediction algorithm in the industry occurs, the following problems are encountered: First, the user's response to advertising is a dynamic process, which will change over time. The prediction algorithm of click-through rate needs to take into account the time factor when applying it to landing.

$$\hat{\rho}_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}, \forall 0 \leq k \leq n \tag{7}$$

$$\hat{\phi}_{kk} = \frac{\hat{D}_k}{\hat{D}}, \forall 0 < k < n \tag{8}$$

Because mobile applications usually use cold start as a startup in order to save memory, most ads have little history in pages, or limited history. In order to solve the historical exposure rate of advertising or new advertising estimates, a hierarchical importance perception factor machine was developed, HIFM provides an effective general framework, as Eq. (9), Eq. (10), the framework combines importance weight and hierarchical learning, after experimental data validation, HIFM better than the existing FMM model in terms of time sensitivity, and the importance of HIFM perception and hierarchical learning plays a significant role in improving the cold start scenario.

$$min\ AIC = nIn\hat{\sigma}_\varepsilon^2 + 2(p+q+1) \tag{9}$$

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-p} + \mu_t + \sum_{j=1}^{q} \theta_j \mu_{t-q} \tag{10}$$

B. The Impact of Information Asymmetry on Real-Time Online Advertising

In order to solve the problem of click rate prediction in industrial applications, a method of screening and combining features with gradient promotion decision tree is proposed by generating a discrete feature vector and taking it as the input parameter of the LR model. GBDT is essentially an integrated learner, as shown in Eq. (11) and Eq. (12), the combination of multiple decision trees can describe the differentiated feature combinations more accurately; compared with the single decision tree, the combination method of GBDT and time series is often applied in the recommendation system.

$$\text{Gini}(D)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2 \tag{11}$$

$$\text{Gini}(D,A)=\frac{|D_1|}{|D|}\text{Gini}(D_1)+\frac{|D_2|}{|D|}\text{Gini}(D_2) \tag{12}$$

Also in the industrial applications, the improvement model is put forward. The model adopts the idea of partition and treatment, group and slice the samples, and LR model is used to predict the partition samples, as shown in Eq. (13) and Eq. (14). Finally, weighted regression is used to combine the results of the partition. The model is proven to fit complex nonlinear functions, and the LS-PLM model using L1 and L2 regulars has good sparsity, which improves the online prediction ability.

$$\Delta D_A=\text{Gini}(D)-\text{Gini}_A(D) \tag{13}$$

$$y=\sum_{k=1}^{K}f_k(x),f_k\subset\Gamma \tag{14}$$

With the in-depth research of deep learning technology in the advertising industry and the breakthrough results, deep neural networks have been realized to fit high-order combination characteristics through nonlinear functions, as shown in Eq. (15) and Eq. (16), so DNN technology is gradually used in the click-through rate prediction model. Based on the DNN model, the deep neural network model of the factorization machine is proposed. This model uses the hidden vector and its weight obtained by FM pre-training as the initial value of the neural network, and then provides the weight of DNN to learn higher-order features, and predicts the click rate of online ads.

$$Obj=\sum_{i=1}^{n}l(y_i,\bar{y}_i)+\sum_{k=1}^{K}\Omega(h_t) \tag{15}$$

$$\Omega(h_t)=\gamma J+\frac{\lambda}{2}\Sigma_{j=1}^{J}\omega_{tj}^2 \tag{16}$$

However, because the DNN model relies on the pre-training of the FM model, it affects the model performance. An improvement is proposed on the DNN model. As shown in Eq. (17) and Eq. (18), the improved PNN model can significantly improve the expression ability of the combined features. The DNN-based model predicts the click rate of online ads by solving the weights of higher-order combined features. We find the importance of low-order features for the prediction of online AD hits and propose Wide and Deep models for the combination of low-order and high-order features.

$$L_t=\sum_{i=1}^{m}L(y_i,f_{t-1}(x_i)+h_t(x_i))+\gamma J+\frac{\lambda}{2}\sum_{j=1}^{J}w_{tj}^2 \tag{17}$$

$$g_{ti}=\frac{\partial L(y_i,f_{t-1}(x_i))}{\partial f_{t-1}(x_i)},h_{ti}=\frac{\partial^2 L(y_i,f_{t-1}(x_i))}{\partial f_{t-1}^2(x_i)} \tag{18}$$

The model is divided into two parts, where the Wide part consists of a generalized linear model, and the Deep part is composed of a DNN model with three hidden layers. In past studies, as shown in Eq. (19) and Eq. (20), click rate prediction and bid optimization are usually carried out in order, firstly minimizing the error between the prediction result and the user response in the real situation by establishing the model. After obtaining the user response prediction result, it is used as input to optimize bids based on other factors such as activity budget, market price, etc.

$$L_t=\sum_{j=1}^{J}[G_{tj}w_{tj}+\frac{1}{2}(H_{tj}+\lambda)w_{tj}^2]+\gamma J \tag{19}$$

$$score=max(score,\frac{1}{2}\frac{G_L^2}{H_L+\lambda}+\frac{1}{2}\frac{G_R^2}{H_R+\lambda}-\frac{1}{2}\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}-\lambda) \tag{20}$$

III. Online Advertising Model Based on Reinforcement Learning

A. CNN Incomplete Information Game and Regret Minimization Algorithm

When a user views the web page or opens a new page in the mobile App, one or more RTB AD spots will appear on the page. The reserved AD space on the page will launch an online AD display request to the supplier platform that provides the advertising agent for the website through the pre-written script code [20, 21]. After receiving the online AD request by clicking the AD display on the user page, the supplier platform SSP will send the AD space information and the web context information to ADX, the AD trading platform that performs the online advertising RTB [22, 23]. AD trading platform ADX receives online AD requests and publishes them to the DSP. In this process, the same online AD request may be sent to multiple different DSPs simultaneously. Fig. 1 shows an overview of the deep reinforcement learning framework. Each demand-side platform publishes details of online advertising requests to advertisers and carries out the first round of online advertising within the demand-side platform. Advertisers can query users' basic personal information, such as education, occupation, gender, etc. After query information, make a decision whether to participate in online advertising auction according to the established online advertising strategy [24, 25]. If you participate in online advertising, you will return your bid to the DSP. After the demand-side platform DSP receives the bid request of each advertiser, it will first rank the bid of each advertiser from the highest to the lowest level.

The advertisers with the highest bids will win the first round of online advertising. The DSP returns the graphic or video information of the top-ranked advertisers together with the bidding price determined by the broad second-highest price mechanism. Each demand side carries the highest-ranked advertising information and bidding price to participate in the second round of online advertising in ADX [26, 27]. ADX ranks

the bids of each demand side, and the highest-ranked demand side wins the second auction. Table I shows the values of AIC, BIC and HQIC, which won the first round of auction by the demand side. The advertising trading platform sends the information of the advertisement to the supplier platform. The supplier platform SSP transmits the advertising text, advertising links or advertising video information of the online advertising winner among the advertisers to the page viewed by the user, and the data will be displayed to the user after rendering through the page [28, 29]. After seeing the advertisement, users may click on the advertisement because of their own interests or be attracted by low discounts, or log in to register the website in the advertisement to complete the transformation, or they may feel that the content of the advertisement is novel and interesting, so they share the advertisement [30]. The complete RTB process describes a two-stage auction. That is, in each DSP internal, the advertisers carry out online advertising auctions. The green box mark part is the second stage of the auction, where each DSP with the highest ads in the first round of online advertising ranking, the second round of online advertising in ADX, competing for the display opportunity of advertising.

Online advertising with accurate positioning of users, and efficient online advertising mechanism, significantly improve users' use experience and delivery efficiency. A top DSP company such as Byte Dance can handle Cookie data from more than 570 million Internet users and use 3,155 attribute tags to represent each Cookie. The DSP sells more than three billion AD displays a day, with each AD display being auctioned off within 50ms. With this Cookies-based audience targeting technology, the market efficiency and effectiveness of RTB advertising have been increased by 50%. It is the Internet big data analysis technology that makes RTB advertising more accurate, controllable and efficient, and also makes RTB become the standard business model of the future online advertising market. The link of RTB is very complex and has a lot of uncertain factors. Fig. 2 shows the updated flow chart of the state-action value function. Therefore, it has certain practical significance to design an applicable online advertising model. When the settlement is CPM, the risk of advertising is estimated and controlled by the demand side. At the same time, because there is no quantification of the user's subsequent behavior, it is difficult to calculate and analyze the advertising effect. When the settlement is made by CPC, the supplier platform SSP can obtain a relatively accurate click rate estimate through the historical user data, and since the subsequent conversion is conducted in the AD demand side site, the AD demand side can make more accurate click value estimation. When the settlement is made by CPA, there is no risk of loss to the demand side, and the supply side is more difficult to operate. So now, in the form of CPA the settlement of online advertising, is gradually decreasing. Through comparison and synthesis, the settlement method of settlement by the click of advertising is the most beneficial to give full play to the advantages of the advertising supply side and the advertising demand side, so the CPC settlement method is widely recognized and accepted in the advertising market. From the perspective of CPC settlement, the online advertising model is discussed.

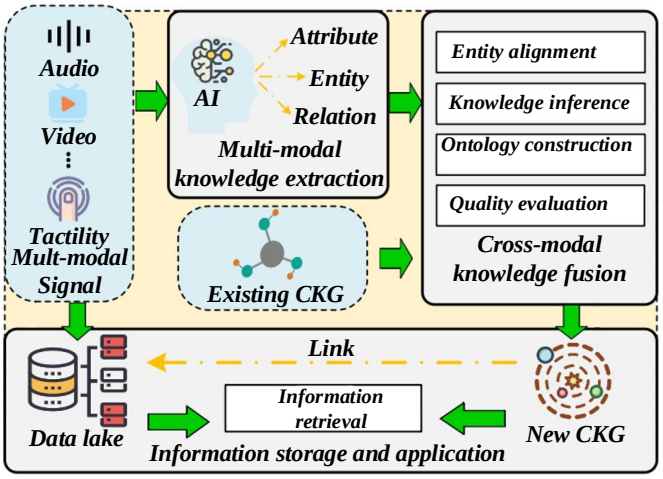

Fig. 1. Overview of the deep reinforcement learning framework.

TABLE I. THE VALUES OF INFORMATION CRITERIA

Information criteria	Price
Aic	1579.70254
Bic	1602.10028
Hqic	1588.73045

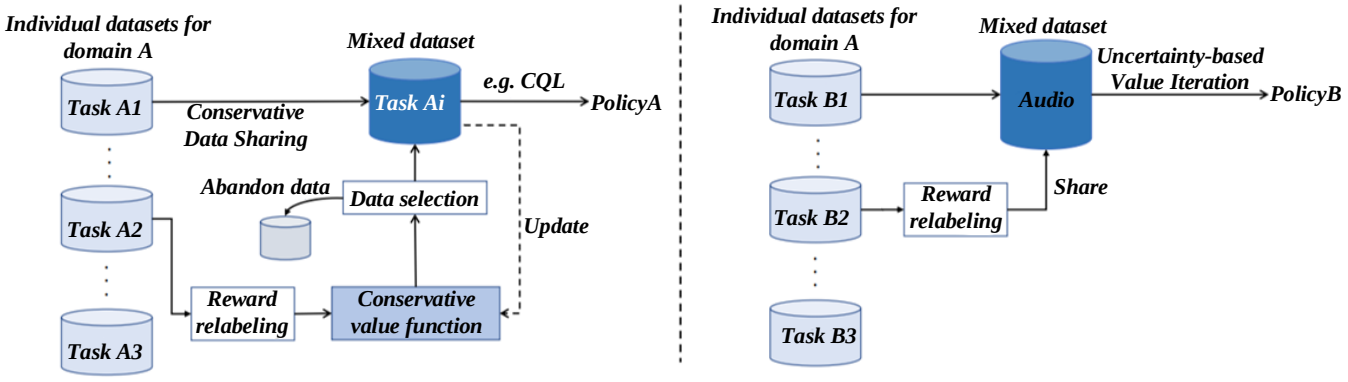

Fig. 2. State-update flow chart of the action value function.

B. Improve the DQN Customer Behavior Characteristic Analysis Model

After receiving the online advertising notification, advertisers use the known information to predict the click-through rate. After the advertiser gets the forecast result, the bid decision is made based on the value that the user's click brings to the brand, the advertising budget, the importance of the advertising space and other information. In most cases, the advertisers' bids are independent of each other. After receiving the quotation returned by each advertiser, DSP will also use known information to predict shot advertisements, calculate eCPC and rank online ads. In this calculation, advertisers with low click-through rate predictions cannot get advertising opportunities with high offers alone. To verify the proposed model, this paper introduces the famous advertising company, the published data training set for the global RTB algorithm competition. After pre-processing the data, it was found that the exposure of the advertisement changed periodically, and the exposure during the weekend period was significantly higher than the weekday's period. Fig. 3 is a graph of feature selection and network structure optimization algorithm. At the same time, the exposure at 12 and 22 points during the rest time is significantly higher than the working and sleep time. Before modeling the time series, the original sequence is required to verify the stationarity requirements.

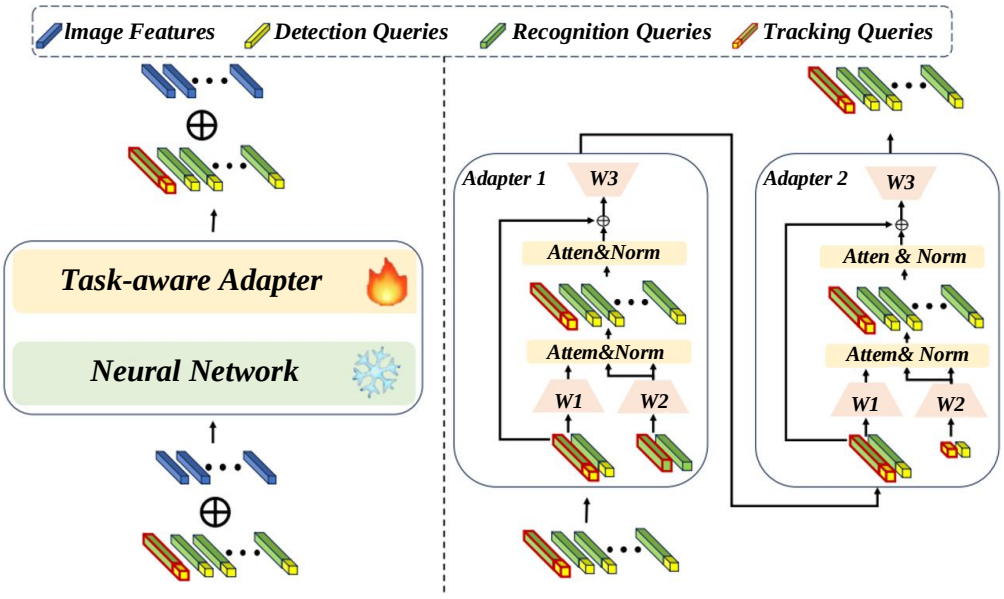

Fig. 3. Feature selection and network structure optimization algorithm.

The pattern of the characteristics of non-stationary sequences at various time points change randomly, which needs to be transformed by differential or logarithmic processing, and then modeling analysis. The randomness test is also known as a white noise test. In general, a white noise test of the stationary sequence is required before modeling to determine whether it has an analytical value. Table II shows the cross-entropy of different algorithms. In the time series model, the more unknown parameters contained, the more the independent variables are included, and the more flexible the corresponding model changes, the higher the accuracy of fitting the model and the greater the likelihood function value; however, when the number of unknown parameters in the model increases, the instability of the model becomes more difficult to fit the model. In general, a time series model is a good model when it considers the fitting accuracy and the number of unknown parameters.

TABLE II. CROSS-ENTROPY OF THE DIFFERENT ALGORITHMS

Algorithm Name	Logloss Cross Entropy
LR	0.423
XGBoost	0.414
ARMA-XGBoost	0.391

In the AD click rate prediction, the classification regression tree is generally chosen as the weak learners. The nature of the classification tree: First, make the basic assumptions, assume that the model is a basic binary tree, and then divide the nodes in the tree by constantly learning the features in the data set, and finally generate the classification tree that meets the expectations. In this article, By dividing it from the top down, The CART classification tree construction using a greedy strategy, Each child node in the tree is split according to the impurity of the subset elements, The urity of the set D that requires training using the Gini Expo measure, For the dataset D included in each node, XGBoost The model training goal is to computational solve the model to find the most appropriate division criterion and the final division value, Making the absolute value of the Gini index difference before and after the split, Fig. 4 shows a comparative evaluation chart of the effect of advertising channels, DA represents the size of the information gain: XGBoost has made great improvements in the algorithm and engineering application of GBDT. GBDT iterates on a set of weak learners, such as a decision tree, and outputs the final prediction results. Specific can be described as the error rate of the iterative decision tree to update the weight of the training set, and through the result of multiple decision tree accumulation as the prediction results output, through the

precise algorithm for all the information gain calculation value, select the maximum feature again using accurate algorithm for segmentation, get the corresponding training feature barrel. However, this order is not the optimal solution. According to Bayesian decision theory, the learning goal of the user response model should be determined by the final bid utility. In the paper, it is proposed that the accuracy of click rate required to predict and the computing resources are not the same within the range of all advertisements to be predicted. The prediction methods and the allocation of computing power should focus on cases with higher return on investment and learn how to predict more accurately. Therefore, the market price and competitive performance are included in the model, and if the investment return of an advertising space is relatively high, the confidence of whether the advertiser can take the advertisement will be predicted. When the confidence is low, the optimization method of click-through rate prediction and the allocated computing resources should be more concentrated than the advertising space with low investment return on ratio.

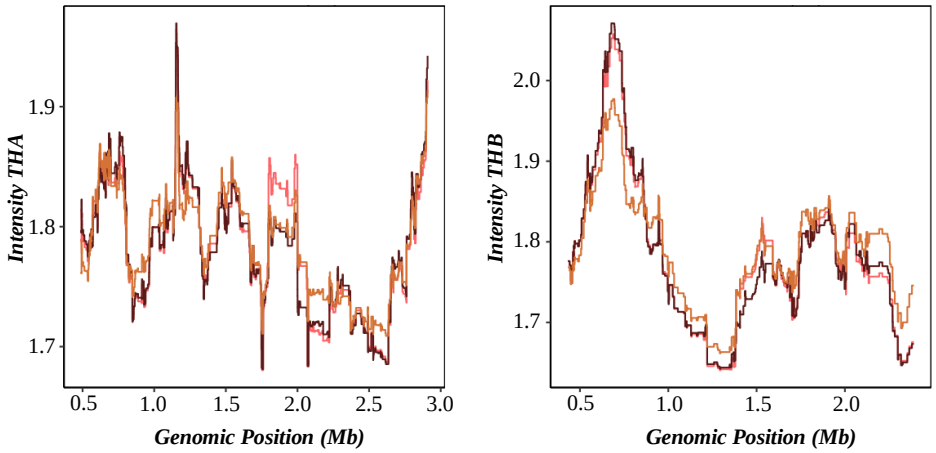

Fig. 4. Comparison and evaluation diagram of the effect of advertising channels.

IV. Research on Customer Behavior Characteristics Analysis and Online Advertising Optimization Based on Deep Reinforcement Learning

Grid search is a model hyperparameter optimization technique, which is essentially an exhaustive method. For each hyperparameter to be determined, a smaller finite set is chosen to explore. Then, we find the Cartesian products of the selected parameters to obtain several combinations of the hyperparameters. The grid search method trains the model using a combination of hyperparameters and selects the combination of hyperparameters that minimize the validation set error as the final parameter of the model. In the process of practical model training, the cross-validation method and the grid search method are usually combined, as the method of parameter evaluation, and this comprehensive method is recorded as the cross-validation grid search method. Fig. 5 shows the time evaluation chart of customer click behavior, and divides the labeled training data in the data set into n-folds for cross-validation. First in the super parameter grid search parameter calculation, and then each set of super parameters into the model for n fold training, select the score of the highest super parameter combination into model, using the model to train the training set data, while using the validation set data validation model training results, get the final results. In this paper, the area enclosed between the receiver working characteristic curve and the coordinate axis and the cross-entropy Logistic Loss is used to evaluate the prediction results of the click rate prediction model. AUC value evaluation index. In the online advertising auction, in order to ensure the maximum benefit and disturb users as little as possible, the accuracy index of the model is very important.

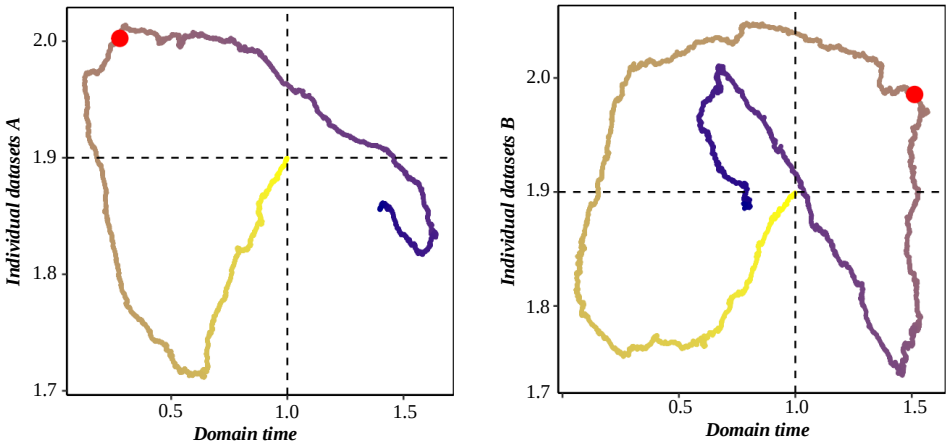

Fig. 5. Time assessment chart of customer click behavior.

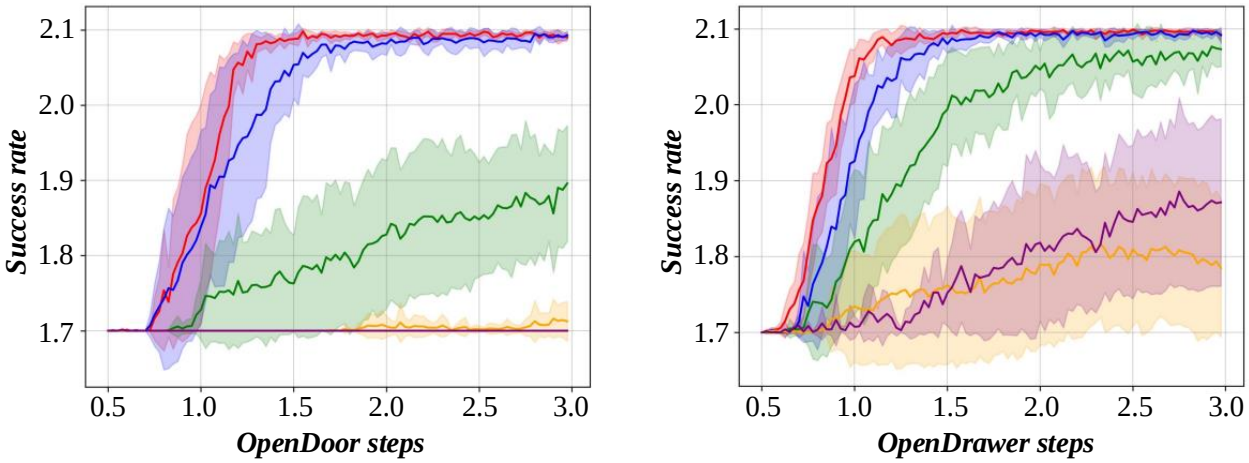

Fig. 6. Assessment chart of AD exposure and conversion rate.

The AUC value is the area obtained by calculus calculation of the ROC curve and the coordinate axis. It indicates the probability that the two samples are randomly selected and predicts the probability that the click rate. Where Rins, i represents the serial number of the sample, M represents the number of positive cases, and N represents the number of counterexamples. AUC value range between [0, 1], when the model AUC value of 0.5 represents a random classifier, the selection of threshold. The higher the AUC score indicates that the better the classifier predicts in the AD click rate prediction problem, the more meaningful the online advertisement. Game theory is based on rigorous mathematical models and introduces the limits of confrontational conflict in the real world. Fig. 6 shows the evaluation chart of advertising exposure and conversion rate, which defines the dominant strategy as follows: assuming that during the game process, the participants will adopt a certain strategy for any strategy choice of the other party. On this basis, if the strategy combination of all the players in the game is the dominant strategy for the other side, then the strategy combination of the two players is called the Nash equilibrium. The RTB online advertising problem discussed in this paper constitutes information asymmetry because the two parties do not know the other party's prediction of the click-through rate. But traditional game theory requires a complete set of situation information and strategies, such as in Texas Hold 'em, the opponent's cards must be a subset of the complete cards; in Go, the opponent's next strategy must be one of the feasible positions on the board. Therefore, the RTB online advertising problem discussed in this paper does not satisfy the incomplete information game; but it can make references for the incomplete information and it's solving methods.

In actual online advertising, the demand parties need to compete with each other in order to win the first auction round. In addition, in the process of continuous auction, both parties can push back the asymmetric information about the click rate by observing the auction results and testing the bid many times. Thus, it develops in the direction of an evolutionary game and achieves the equilibrium of the game through trial and error. In the first round of an auction, different demand parties may offer different bids, but they all face the DSP when predicting the click rate to calculate the eCPC ranking. Therefore, the scenario of the first round of auction is simplified, considering only the interaction between a single AD demand side and the DSP, and not the competition between the demand sides. In the process of continuous auction, the AD demand side and DSP may speculate positively or reverse the information about each other to achieve the dynamic equilibrium of evolution. However, in the early stage of the auction, designing the bid model for the unequal information between the parties is still beneficial to gain more benefits, and helps to seize the first advantage in evolution. In recent years, the ability to process large-scale data, discover the underlying features and extract the underlying features, so as to achieve specific goals more accurately. Fig. 7 shows the evaluation diagram of customer interest preferences at different ages. As a learning method with interactive ability, the online advertising decision model based on reinforcement learning is essentially a Markov decision process. Markov decision process has the characteristics of interacting with the environment, so this kind of model has the natural characteristics of modeling online advertising decision behavior, and the application of exploration mechanism can make the agent more fully explore the state and action space, and improve the accuracy and diversity of decision results to a certain extent. Among them, the online advertising decision system based on reinforcement learning consists of the current environment of the agent, the budget, the remaining advertising space and the information mastered by the DSP. In the process of the interaction between the agent and the environment, the decision agent constantly explores the decision scheme for the agent to obtain the maximum revenue according to the given budget, obtained traffic and other information, decide the bid at each auction time, and finally spends the budget to presents a decision scheme.

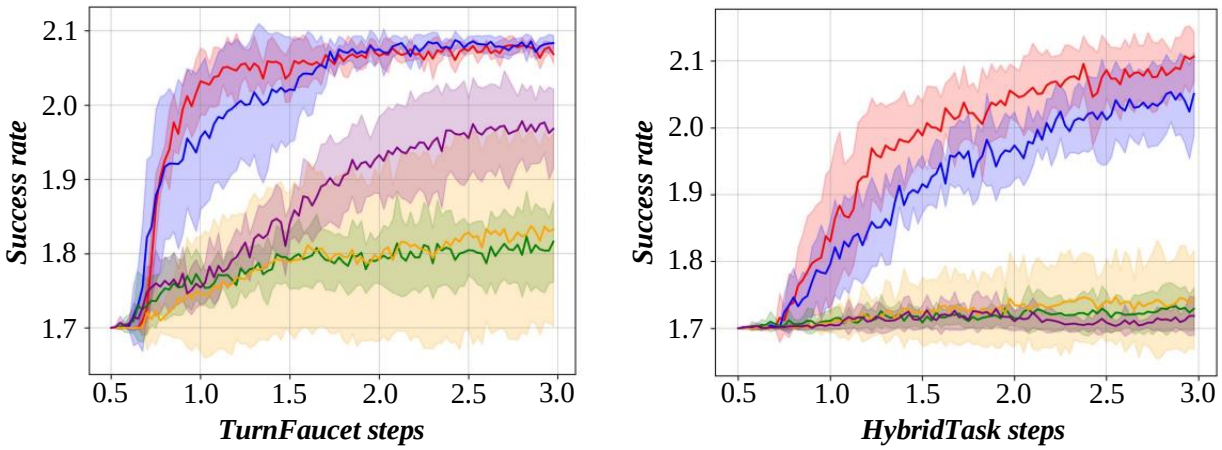

Fig. 7. Assessment chart of customer interest preferences at different ages.

Action space A: Action A represents the current moment, when the decision agent decides for the probability that the decision agent takes action a understate s, updates the budget and click rate, and moves to the next state S '. If the AD is won; if the AD is not won, one online advertising opportunity is lost. The decision agent can be converted to an immediate reward *R (s, a)* based on the feedback results. Discount factory: the factor that determines the value of long-term rewards at the current moment, the decision agent uses the greedy method, which is only considered for the immediate rewards; when y=1, the subsequent rewards have the same value at the current moment. Strategy π represents the basis for an individual taking an action, with its data described as a conditional probability distribution, namely the probability of taking an action *a* at state *s*. Action value function: the decision agent takes action A according to the strategy. The value function is an expectation function. Although Rt + 1, if only referring to the delay reward, it is easy to ignore the global situation and fall into the local optimal solution. Therefore, it is necessary to consider the delay reward of the current action and the potential delayed reward of the subsequent action. This section refers to the CFR algorithm for solving incomplete information games and improves the setting of the reward function in DQN. The regret value calculation method in CFR is introduced to set the reward function by taking the bid action and getting the environmental feedback regret value. The traditional Q-Learning reinforcement learning algorithm uses Q table to store data when making online advertising decisions. However, when the bidding environment is complex and the advertising space becomes more, Q table becomes huge and causes storage problems, and the search problem caused by the increase of data volume is easy to lead to the explosion of algorithm dimension. At present, in order to solve this problem, the industry generally uses other solutions to replace the Q value table, the most widely used is to use function approximation to replace. However, because the function approximation is calculated by calculating the value function, this alternative method is prone to the instability of the algorithm model and the failure to converge. The DQN algorithm combines the advantages of the traditional reinforcement learning algorithm Q-Learning and the deep neural network, which significantly improves the instability problem when the algorithm approximates the value function, which solves the problem of model instability and convergence to a certain extent.

V. Experimental Analysis

The core idea of the DQN algorithm is experience playback. During the training process, the reward results and model state update of each environment interaction are saved to the specified position, so that they can be used for the subsequent update operation of the target Q value. Fig. 8 for customer purchase path evaluation diagram, generally speaking, through the Q network calculated Q value and through experience playback target Q value will have some error, when the need to reduce the error can through the Q value gradient backpropagation to update the deep neural network parameter w, when the parameter w convergence, can get less error of approximate Q value.

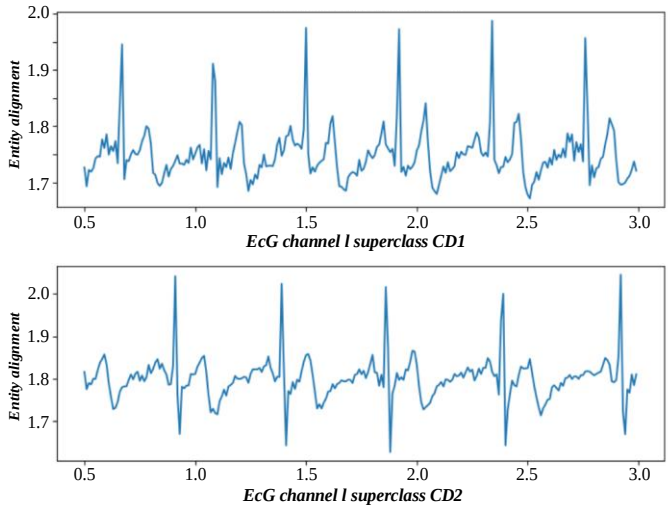

Fig. 8. Customer purchase path evaluation chart.

In reinforcement learning, the agent needs to explore in the current state to make better decisions, so as to obtain better returns, and choose the corresponding strategy to make the expected decision accordingly. The commonly used action selection strategies are the e-greedy algorithm. This paper selects the e-greedy algorithm. E-Greedy algorithm is a multi-arm gambling algorithm improved based on the greedy algorithm. Fig. 9 shows the evaluation diagram of the correlation of advertising creative type and click-through rate. In order to

avoid the agent always choosing the action with the current highest return to make the calculation probability of action execution. To obtain a better selection strategy, the strategy is randomly selected at probability E in the initial state.

In order to verify and analyze the prediction model and the subsequent online advertising model, the data set published by Tencent's 2019 advertising algorithm competition was selected. Fig. 10 is the evaluation chart of advertisement exposure during active time, which will hereinafter referred to as Tencent Data Set. In RTB transactions, Tencent can act as an advertiser to promote their games, or as an advertiser to publish ads for other brands on social platforms.

Therefore, Tencent's data set includes the whole process data and log of online advertising from online advertising to display and complete transformation: including historical exposure log, user characteristic and attribute data, advertising static data and advertising operation data. According to the assumptions of the model, the information inequality between the DSP and the demand side is discussed separately. Fig. 11 shows the cost-benefit analysis and evaluation chart of advertising, considering the following three experimental designs: the click rate predicted by the preset DSP is higher than the demand side, the click rate predicted by the preset DSP is lower than the forecast result of the demand side; and the real prediction results of the participants after dividing the data set. In the ideal stage of the initial transaction, by analyzing the winning results feedback by DSP, the estimated range of the DSP click rate is obtained.

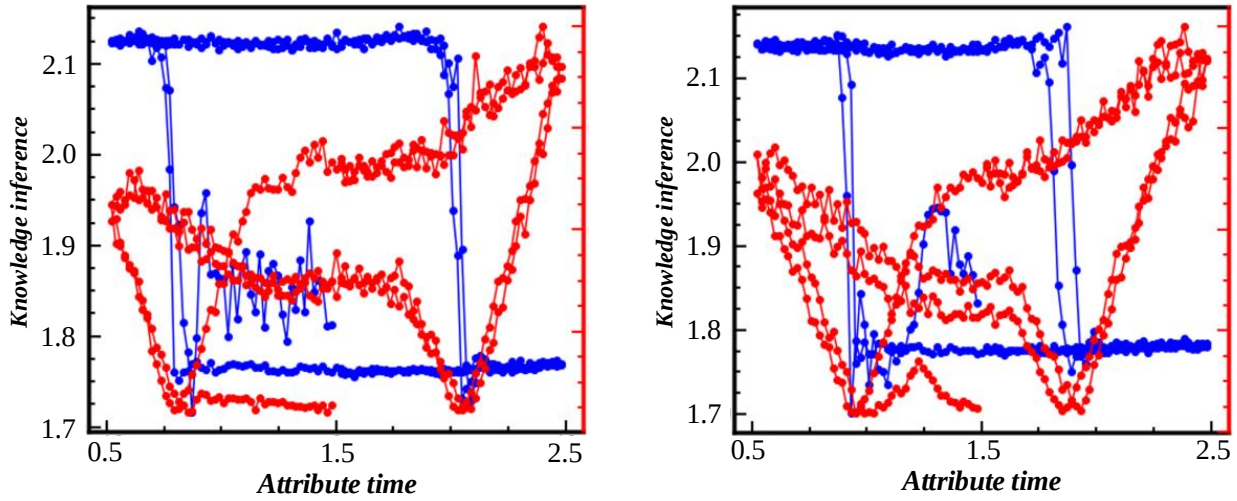

Fig. 9. Evaluation diagram of the correlation between advertising creative type and click-through rate.

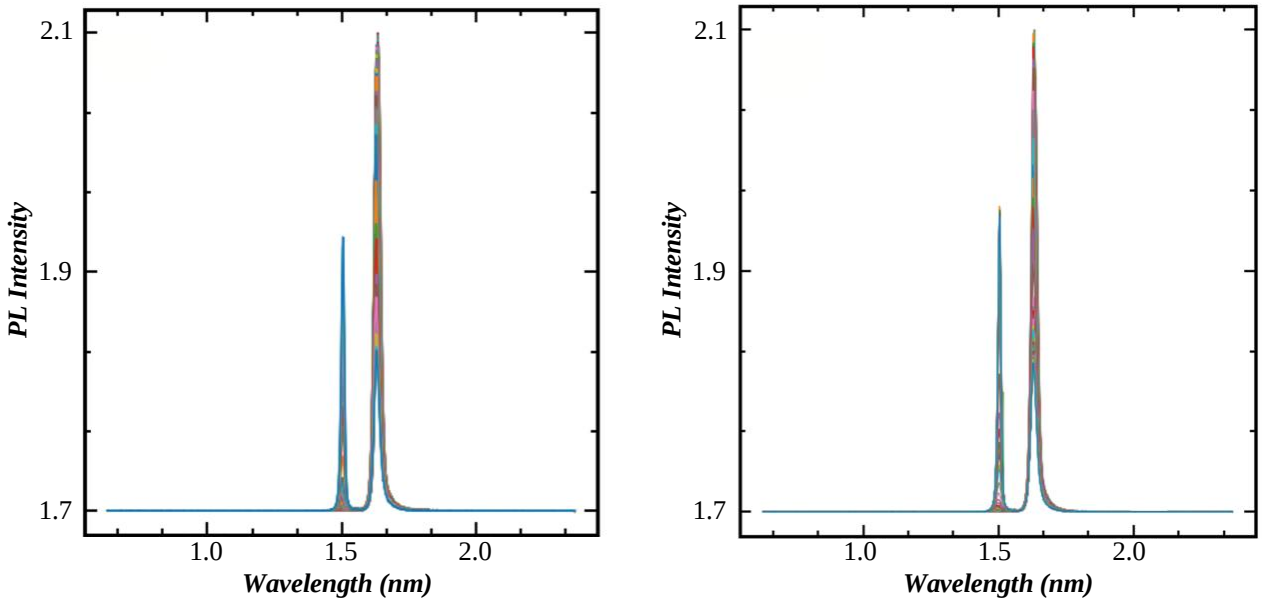

Fig. 10. Evaluation chart of advertising exposure during user active time.

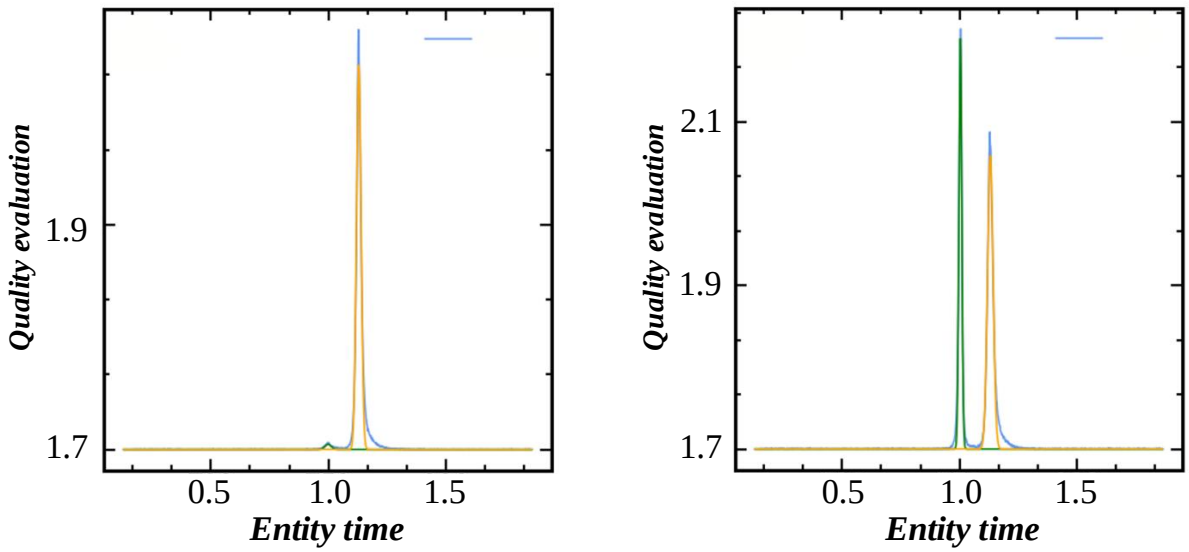

Fig. 11. Cost-benefit analysis and evaluation chart of advertising delivery.

VI. Conclusion

This paper introduces the common settlement method of online advertising, and introduces the eCPC formula of calculating online advertising ranking in the real production environment. Public information and private information in RTB transactions are defined and divided. This paper analyzes the problems caused by information asymmetry in the process of online advertising. Considering the difference in DSP and advertisers in the first auction, there is information asymmetry in online advertising. In this paper, the reinforcement learning model is designed, and the AD click rate is underestimated or overestimated by DSP, and the real prediction situation is designed respectively. The experiment shows that in the three scenarios, the click rate of improved DQN is higher than the traditional online advertising model. In the case where the click-through rate of advertising is overestimated by DSP, the high-frequency bid low-price advertising can maximize the interests of advertisers; in the situation where the click-through rate is underestimated by DSP, the low frequency for high-value advertising space is conducive to the exposure of the advertisement; in real scenes, using 10% of the budget to explore the environment, the price distribution of the final bid of the decision agent is the same as the distribution of click-through rate prediction results.

DSP and advertisers' click-through rate estimates follow a random distribution, and the prediction results are different in different ads. The prediction model of the final DSP has 23 inputs for training features and advertisers have 21 inputs for training features. According to the prediction results, the gap of DSP and advertisers within 0.02 was 15.7%, DSP higher than advertisers 41.7%, and DSP lower than advertisers 42.6%. LinBid Online advertising of Experiment 1 and 2, the LinBid winning rate always increases linearly with the budget. In contrast, the winning rate of the proposed improved DQN algorithm fluctuates greatly with the budget, and the improved DQN algorithm increases significantly when the budget increases. If the DSP predicted click-through rate is higher than the advertisers, the advertisers adopt the "positive" bidding strategy, high frequency bid low price ads, and achieve the click target; if the DSP predicted click-through rate is lower than the advertisers, the advertisers adopt the "cautious" bidding strategy, low frequency bid high price advertising, high price advertising, can significantly improve the return-on-investment ratio. When DSP has accumulated the user information of the brand and the transaction environment is gradually complex, DSP may overestimate or underestimate the click-through rate prediction results of advertisers: when the budget is low, advertisers can consider using the relatively stable LinBid online advertising model to bid based on the historical transaction price and the clicks predicted by advertisers. When the budget is high, and the pursuit of high profit, 10% of the budget to explore the proportion of the number of low-priced and high-priced ads in the final bid follows the proportion of the number of ads whose click rate is overestimated and undervalued by DSP.

References

[1] Abadi, Z. J. K., Mansouri, N., & Javidi, M. M. (2024). Deep reinforcement learning-based scheduling in distributed systems: a critical review. Knowledge and Information Systems, 74.

[2] Abdulazeez, D. H., & Askar, S. K. (2023). Offloading Mechanisms Based on Reinforcement Learning and Deep Learning Algorithms in the Fog Computing Environment. Ieee Access, 11, 12554-12585.

[3] Alipio, M., & Bures, M. (2023). Deep Reinforcement Learning Perspectives on Improving Reliable Transmissions in IoT Networks: Problem Formulation, Parameter Choices, Challenges, and Future Directions. Internet of Things, 23, 20.

[4] Allaoui, T., Gasmi, K., & Ezzedine, T. (2024). Reinforcement learning based task offloading of IoT applications in fog computing: algorithms and optimization techniques. Cluster Computing-the Journal of Networks Software Tools and Applications, 26.

[5] Almazrouei, K., Kamel, I., & Rabie, T. (2023). Dynamic Obstacle Avoidance and Path Planning through Reinforcement Learning. Applied Sciences-Basel, 13(14), 20.

[6] Chung, J. H., Fayyad, J., Al Younes, Y., & Najjaran, H. (2024). Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding. Artificial Intelligence Review, 57(2), 36.

[7] Delgado, J. M. D., & Oyedele, L. (2022). Robotics in construction: A critical review of the reinforcement learning and imitation learning paradigms. Advanced Engineering Informatics, 54, 24.

[8] Dong, L., He, Z. C., Song, C. W., & Sun, C. Y. (2023). A review of mobile robot motion planning methods: from classical motion planning workflows to reinforcement learning-based architectures. Journal of Systems Engineering and Electronics, 34(2), 439-459.

[9] Esteso, A., Peidro, D., Mula, J., & Díaz-Madroñero, M. (2023). Reinforcement learning applied to production planning and control. International Journal of Production Research, 61(16), 5772-5789.

[10] Faria, R. D., Capron, B. D. O., Secchi, A. R., & de Souza, M. B., Jr. (2022). Where Reinforcement Learning Meets Process Control: Review and Guidelines. Processes, 10(11), 31.

[11] Frikha, M. S., Gammar, S. M., Lahmadi, A., & Andrey, L. (2021). Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey. Computer Communications, 178, 98-113.

[12] Gao, Q. H., & Schweidtmann, A. M. (2024). Deep reinforcement learning for process design: Review and perspective. Current Opinion in Chemical Engineering, 44, 10.

[13] Ghotbi, M., & Zahedi, M. (2024). Predicting price trends combining kinetic energy and deep reinforcement learning. Expert Systems with Applications, 244, 12.

[14] Greguric, M., Vujic, M., Alexopoulos, C., & Miletic, M. (2020). Application of Deep Reinforcement Learning in Traffic Signal Control: An Overview and Impact of Open Traffic Data. Applied Sciences-Basel, 10(11), 25.

[15] Gupta, S., Singal, G., & Garg, D. (2021). Deep Reinforcement Learning Techniques in Diversified Domains: A Survey. Archives of Computational Methods in Engineering, 28(7), 4715-4754.

[16] Han, D., Mulyana, B., Stankovic, V., & Cheng, S. (2023). A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation. Sensors, 23(7), 35.

[17] Hasan, Z., & Roy, N. (2021). Trending machine learning models in cyber-physical building environment: A survey. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 11(5), 13.

[18] Hickling, T., Zenati, A., Aouf, N., & Spencer, P. (2024). Explainability in Deep Reinforcement Learning: A Review into Current Methods and Applications. Acm Computing Surveys, 56(5), 35.

[19] Hou, H. H., Jawaddi, S. N. A., & Ismail, A. (2024). Energy efficient task scheduling based on deep reinforcement learning in cloud environment: A specialized review. Future Generation Computer Systems-the International Journal of Escience, 151, 214-231.

[20] Hua, J., Zeng, L. C., Li, G. F., & Ju, Z. J. (2021). Learning for a Robot: Deep Reinforcement Learning, Imitation Learning, Transfer Learning. Sensors, 21(4), 21.

[21] Ibrahim, A. M., Yau, K. L. A., Chong, Y. W., & Wu, C. (2021). Applications of Multi-Agent Deep Reinforcement Learning: Models and Algorithms. Applied Sciences-Basel, 11(22), 40.

[22] Jogunola, O., Adebisi, B., Ikpehai, A., Popoola, S. I., Gui, G., Gacanin, H., & Ci, S. (2021). Consensus Algorithms and Deep Reinforcement Learning in Energy Market: A Review. Ieee Internet of Things Journal, 8(6), 4211-4227.

[23] Ju, H., Juan, R. S., Gomez, R., Nakamura, K., & Li, G. L. (2022). Transferring policy of deep reinforcement learning from simulation to reality for robotics. Nature Machine Intelligence, 4(12), 1077-1087.

[24] Khoei, T. T., Slimane, H. O., & Kaabouch, N. (2023). Deep learning: systematic review, models, challenges, and research directions. Neural Computing & Applications, 35(31), 23103-23124.

[25] Li, C. X., Zheng, P., Yin, Y., Wang, B. C., & Wang, L. H. (2023). Deep reinforcement learning in smart manufacturing: A review and prospects. Cirp Journal of Manufacturing Science and Technology, 40, 75-101.

[26] Lin, B. H. (2024). Reinforcement learning and bandits for speech and language processing: Tutorial, review and outlook. Expert Systems with Applications, 238, 32.

[27] Massaoudi, M., Chihi, I., Abu-Rub, H., Refaat, S. S., & Oueslati, F. S. (2021). Convergence of Photovoltaic Power Forecasting and Deep Learning: State-of-Art Review. Ieee Access, 9, 136593-136615.

[28] Massaoudi, M. S., Abu-Rub, H., & Ghrayeb, A. (2023). Navigating the Landscape of Deep Reinforcement Learning for Power System Stability Control: A Review. Ieee Access, 11, 134298-134317.

[29] Mohammed, M. Q., Chung, K. L., & Chyi, C. S. (2020). Review of Deep Reinforcement Learning-Based Object Grasping: Techniques, Open Challenges, and Recommendations. Ieee Access, 8, 178450-178481.

[30] Munikoti, S., Agarwal, D., Das, L., Halappanavar, M., & Natarajan, B. (2023). Challenges and Opportunities in Deep Reinforcement Learning With Graph Neural Networks: A Comprehensive Review of Algorithms and Applications. Ieee Transactions on Neural Networks and Learning Systems, 21.