Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Optimization of Civic Education Teaching Strategies Based on Reinforcement Learning and Its Numerical Simulation

Wei Zheng[1], Qinghua Lu[2,] ✉

[1] *Student Affairs Office, Hunan Railway profession College, Zhuzhou, Hunan, 412000, China*
[2] *School of Marxism, Hunan Railway profession College, Zhuzhou, Hunan, 412000, China*

ABSTRACT
Driven by the core qualities of the Civics discipline, the requirements of curriculum reform and the needs of teaching practice, the optimization of teaching strategies has become particularly urgent in the field of Civics education. The article introduces the Markov decision-making process and basic elements of reinforcement learning, combines the Q learning algorithm with neural networks, and constructs a deep reinforcement learning model (IDQN) for multiple intelligences with collaborative scheduling. Based on this, a numerical simulation experiment of deep reinforcement learning strategy in Civics teaching was designed and implemented. Through experimental analysis: when the recommended path is 30, the IDQN model has the best learning path recommendation effect, with an IKL of 0.477. The model also has excellent performance in the allocation of teaching resources, with the accuracy, recall and F1 value of 5 tests above 90%. After the numerical simulation of Civic Education teaching, the learning interest, attitude, and motivation of students in the experimental group increased by 27.52% to 34.49%. Under this influence, combined with the learning path and resource allocation provided by the IDQN model, students in the experimental group showed a significant improvement in their learning effect, and the average score of Civic Education Theory was 6.06 points higher than that of the control group.

Keywords: Deep learning, Markov decision making, Q-learning, IDQN, numerical simulation, Civic education teaching

1. Introduction

Civic and political education is an important part of our country's education and teaching system, and it is necessary to educate and teach students at any stage of any grade. The educational content of Civic and political education is more diverse, including ideological character, political literacy, scientific thinking, etc., all belong to the components of Civic and political education [1]. However, the current

✉ Corresponding author.
E-mail address: hntd_zw@sina.com (Q. Lu).
Received 10 February 2024; Revised 25 May 2024; Accepted 15 October 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-049

teaching strategy of Civic and Political Education can not meet the needs of contemporary students, how to develop an effective teaching strategy has become an urgent problem, and reinforcement learning can contribute to this.

Reinforcement learning is a method of machine learning that aims to enable intelligences to learn how to make optimal decisions by interacting with the environment to achieve specific goals [2-3]. Unlike supervised and unsupervised learning, reinforcement learning is a reward-based learning approach that guides intelligences through reward signals [4-5]. In reinforcement learning, an intelligent body interacts with the environment by mapping the observed information to a specific action, and the environment will provide feedback on whether the intelligent body's action is correct and reward or penalize it accordingly [6]. The goal of the intelligent body is to maximize the sum of long-term rewards by learning the optimal strategy. In reinforcement learning, commonly used algorithms include Q-learning and deep reinforcement learning.Q-learning is a table-based method that updates the Q-value function by using a greedy or ε-greedy policy and gradually converges to the optimal value function [7-8]. Deep reinforcement learning combines the ideas of deep and reinforcement learning by using a neural network to approximate the value function and policy and optimizing the network parameters through a backpropagation algorithm [9-10]. An important issue in reinforcement learning algorithms is the balance between exploration and exploitation. Exploration refers to trying unknown states and actions in order to obtain more information about the environment; exploitation refers to using known strategies to maximize rewards [11-12]. To balance exploration and utilization, an ε-greedy strategy is often used, where exploration is chosen with a certain probability and utilization is chosen with the rest of the probabilities [13]. Reinforcement learning has a wide range of applications in many fields, such as education field, art field, machine control, and automatic driving [14-17]. Therefore, by building appropriate mathematical models and using corresponding algorithms, reinforcement learning can solve many complex problems, such as applied to the teaching of ideological education.

The study explores the application of deep reinforcement learning in the optimization of teaching strategies in Civic and Political Education, and conducts empirical research in combination with teaching numerical simulation. The urgency of the optimization of teaching strategies in Civic and Political Education is clarified, and the importance of accurately locating students' needs is emphasized. In terms of teaching numerical simulation, the study builds a multi-intelligence body IDQN algorithm based on the DQN deep reinforcement learning algorithm to simulate and optimize the optimization strategy of Civic and Political Education. By constructing teaching environment states and behavioral strategies, the algorithm can automatically learn so as to provide personalized teaching suggestions and coaching for Civics teaching.

2. The urgency of optimizing teaching strategies for Civic Education

Civic and political education is a course that is constantly updated with the times, with distinctive characteristics of the times and Chinese characteristics. In recent years, with the development of Civic and Political Education, new teaching methods such as big concept teaching and big unit teaching have emerged in response to the trend, and both big concept and big unit teaching are designed to help students strengthen the knowledge connection, get rid of the superficial learning state, deepen the understanding of knowledge, realize the transfer of the application of political knowledge, and form the core literacy. From another point of view, the big concept and big unit teaching is actually to help students realize the depth of learning of Civic and Political Education.

2.1. Disciplinary core literacy promotion

The era of information globalization has raised higher requirements for the training of human resources, and has given rise to the buzzword of “core literacy”, which is an educational term that is

adapted to the needs of social development and to the needs of one's own lifelong development. The core literacy of the discipline is the good literacy formed by students in terms of values, character, and ability, which is the study of "what kind of people to cultivate". The Civic Education Curriculum Standards (2017 Edition Revised in 2020) clearly states that teachers should change their teaching methods in order to cultivate students' core literacy, which requires teachers to give full play to the auxiliary role of enhanced learning in the teaching of Civic Education.

2.2. Curriculum reform requirements

Compared with the old textbooks, the new teaching materials of Civic and Political Education are more in line with the students' life reality, and the overly theoretical knowledge has been diluted, which requires teachers to optimize their teaching strategies. Civic and political education is a comprehensive and active subject course, which requires teachers of Civic and political education to optimize their teaching strategies and teach students the methods of learning, and only when students really learn and believe can they really understand and use them. Strengthening learning is an important way to promote the deep development of the curriculum and teaching reform.

2.3. Teaching practice needs

At present, the learning of the Civic and Political Education Program is still in a shallow state, and students stay in a shallow memory of what they have learned, and these problems cannot be solved by repeated lectures and repetitive training alone. Therefore, it is necessary to carry out the teaching practice of intensive learning to change the previous shallow learning state and move towards deep learning.

3. Civic education teaching based on multi-intelligence deep reinforcement learning

Through the above analysis of the urgency of optimization of teaching strategies in Civics education, this section proposes a deep reinforcement learning method based on multiple intelligences to be used as an optimization strategy for Civics education teaching. Deep reinforcement learning is widely used in the personalized teaching of Civic and Political Education, and this paper aims to use the deep reinforcement learning model to analyze the personalized needs of students and provide them with accurate and practical personalized learning paths and teaching resources, so as to optimize the strategies of the traditional teaching methods.

3.1. Enhanced learning

The two subjects in reinforcement learning are the intelligent body and the environment, and the intelligent body is an individual that can perceive the environment and take corresponding actions through the changes of the environment so as to accomplish specific goals. The environment is the space where the intelligent body is located, which can change accordingly with the action chosen by the intelligent body, and the interaction process between the intelligent body and the environment is shown in Figure 1.

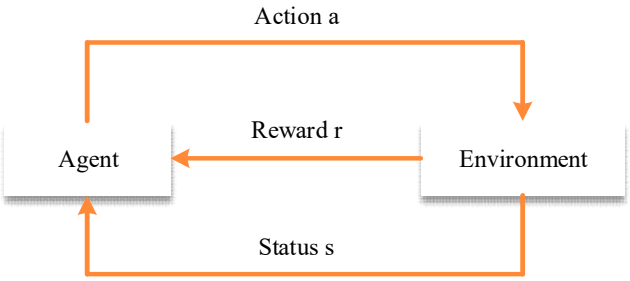


Fig. 1. Intelligent body and environmental interaction process

3.1.1. Markov decision-making process. A Markov Decision Process [18] (MDP) is a decision process in which, given all the states and action selection strategies experienced in the past, the state at the next moment is determined only by the present state and is independent of the past state, as follows:

$$P[s_{t+1} \mid s_t] = P[s_{t+1} \mid s_1, s_2, ..., s_t] \tag{1}$$

A Markov decision process can be represented by a quaternion $< S, A, R, P >$, where the parts are defined as follows:

$S = \{s_0, s_1, ..., s_n\}$ denotes the set consisting of all the states of the intelligent body in the environment, s_p denotes the p th state of the intelligent body in the environment.

$A = \{a_0, a_1, ..., a_m\}$ denotes the set of all actions that can be chosen by the intelligence in the current environment, and a_q denotes the q th action that can be chosen by the intelligence.

$R : S \times A \to R$ denotes the reward function that represents the action selection made by the intelligent body in the environment, $r_t = R(s_t, a_t)$ denotes the reward obtained by the intelligent body for selecting action a_t while in state s_t, and the reward value takes the range of all real numbers.

$P : S \times A \times S \to [0,1]$ denotes the state transfer probability function of the intelligent body in the environment, and $P(s_{t+1} \mid s_t, a_t)$ denotes the probability of the intelligent body's state transitioning to s_{t+1} after selecting action a_t in state s_t, with a range of values between 0 and 1.

3.1.2. Enhanced learning fundamentals. In reinforcement learning, in addition to the intelligences and the environment, there are other components such as strategy π, reward r and value function.

1) Strategy π. In the process of interaction between the intelligent body and the environment, when the intelligent body is in state s_t, it chooses action a_t to reach the next state s_{t+1} according to a certain principle, which is the policy π. The policy can be expressed as a mapping function from the state of the intelligent body to the action:

$$\pi(a \mid s) = P(A_t = a \mid S_t = s) \tag{2}$$

2) Reward r. Reward r is a feedback evaluation that the intelligent body receives after choosing action a_t in state s_t. When the reward is a large positive number, it means that the action chosen by the intelligent body in the current state has a positive effect, and strategy π will be more inclined to choose the action in state s_t.

3) Value function. The cumulative reward G_t at time step t is related to the rewards at multiple future time steps, and it is difficult to accurately measure the future rewards during training with the cumulative reward because the change of strategy during training will make the difference in the future action selection so that all the rewards after that time step will change, which will cause the

cumulative reward to change greatly. The state-value function $V_\pi(s)$ is used to calculate the expected value, which is given in the following formula:

$$V_\pi(s) = E[R_t \mid s_t = s] = E\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s\right] \tag{3}$$

3.1.3. Q-learning algorithm. The Q-learning algorithm [19] learns the state-action value function $Q(s,a)$ by updating it with the rewards obtained when different actions are selected in each state during the interaction between the intelligent body and the environment to obtain the optimal strategy in each state. The updating formula for the Q-value function in the Q-learning algorithm is as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \Delta Q(s_t, a_t) \tag{4}$$

And the update error $\Delta Q(s_t, a_t)$ of the Q-valued function at time step t is given by:

$$\Delta Q(s_t, a_t) \leftarrow r_t + \gamma \max_{a \in A} Q(s_{t+1}, a_t) - Q(s_t, a_t) \tag{5}$$

Thus Q The full form of the learning algorithm update formula is as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a \in A} Q(s_{t+1}, a_t) - Q(s_t, a_t)] \tag{6}$$

3.2. DQN deep reinforcement learning algorithm

The DQN algorithm [20] builds on the Q-learning algorithm in reinforcement learning by using a neural network to fit the Q-value function nonlinearly, fitting the Q-values of discrete state-action pairs to a continuous function.

The DQN algorithm solves the problem of model non-convergence due to data instability during training by incorporating a target network θ^-. $Q(s,a \mid \theta_i)$ denotes the Q-valued function obtained by selecting action a at the ird iteration when the main network is in state s. $Q(s,a \mid \theta_i^-)$ denotes the output of the target network, and θ_i^- is the parameter of the target network at the ith iteration, which can be calculated by the following formula for the Q value of the target network:

$$Y_i = r + \gamma \max_a Q(s', a' \mid \theta_i^-) \tag{7}$$

Wherein, r denotes the reward obtained by selecting an action a in the current state s, s' is the next state entered after selecting an action a in the current state s, and a' denotes all the actions that can be selected in the next state s'. During the neural network training process, the parameters θ of the main network are updated in real time, and the parameters θ of the main network are assigned to the target network θ^- after every certain number of iteration steps N.

The DQN algorithm updates the parameters in the neural network by minimizing the loss function $L(\theta_i)$ between the main network Q value $Q(s,a \mid \theta_i)$ and the target network Q value Y_i. The loss function is given by:

$$L(\theta_i) = E_{s,a,r,s'}[(Y_i - Q(s,a \mid \theta_i))^2] \tag{8}$$

The gradient of the loss function $L(\theta_i)$ is obtained by taking the partial derivative of parameter θ_i in the loss function as follows:

$$\nabla_{\theta_i} L(\theta_i) = E_{s,a,r,s}[(Y_i - Q(s,a \mid \theta_i)) \nabla_{\theta_i} Q(s,a \mid \theta_i)] \tag{9}$$

Finally the main network parameters θ_i are updated using gradient descent:

$$\theta_{i+1} = \theta_i - \alpha \nabla_{\theta_i} L(\theta_i) \tag{10}$$

3.3. Multi-intelligence deep reinforcement learning

Due to the limitation of the application conditions of the single-intelligent body algorithm, it can only be used in a small number of simple scenarios that meet Markov's conditions, but cannot be applied to most real scenarios. Therefore, there is a need to extend the single-intelligent-body algorithm so that it can be applied to multi-intelligent-body environments.

3.3.1. Centralized collaborative scheduling approach. The writing scheduling method [21] used in this paper is centralized, and the interaction process of the centralized collaborative scheduling deep reinforcement learning algorithm is shown in Figure 2. Centralized deep reinforcement learning regards the multi-intelligence system as a superintelligence system with complex network structure, uses a core intelligence to organize and coordinate the superintelligence system, and uses the state and action of the superintelligence to represent the joint state and joint action of the multi-intelligence, respectively.

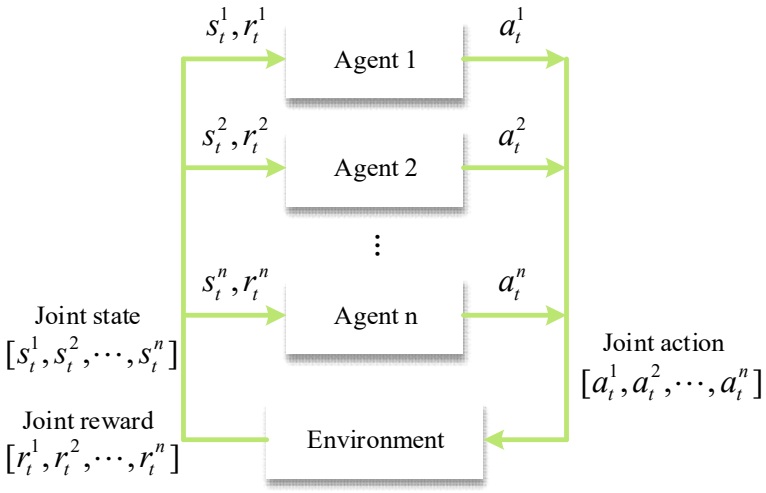


Fig. 2. Interactive process of centralized deep reinforcement learning algorithm

3.3.2. IDQN algorithm. In multi-intelligent body deep reinforcement learning, an independent algorithm is adopted in most cases to assign a separate training network for each intelligent body. The independent deep Q-network (IDQN) algorithm is a standalone deep reinforcement learning algorithm obtained by generalizing the single-intelligent body DQN algorithm to a multi-intelligent body system.The interaction process between the IDQN algorithm and the environment is shown in Figure 3.

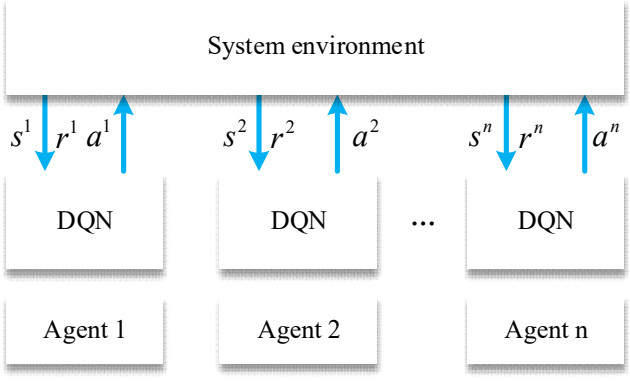


Fig. 3. Interaction process between IDQN algorithm and environment

The IDQN algorithm assumes that the Q-value computations of all the intelligences in the environment do not affect each other, thus decomposing the joint Q-value $Q(s,a)$ of all the intelligences in centralized deep reinforcement learning into the sum of the individual Q-values of the intelligences:

$$Q(s,a)=\sum_{i=1}^{n}Q^{i}(s^{i},a^{i}) \tag{11}$$

The formula for the optimal joint action in centralized deep reinforcement learning is:

$$a^{*}=*\underset{a}{argmax}\,Q(s,a)=*\underset{a^{1},a^{2},...,a^{n}}{argmax}\,Q(s,a^{1},a^{2},...,a^{n}) \tag{12}$$

The solution equation after the decomposition of the optimal joint action is as follows:

$$a^{*}=[\underset{a}{\text{argmax}}\,Q^{1}(s^{1},a),\underset{a}{\text{argmax}}\,Q^{2}(s^{2},a),...,\underset{a}{\text{argmax}}\,Q^{n}(s^{n},a),] \tag{13}$$

4. Implementation of intensive learning under numerical simulation in the teaching of Civics

4.1. Purpose of the experiment

The purpose of this experimental study is to explore whether deep reinforcement learning based on numerical simulation is helpful for improving the teaching effect of Civic and Political Education, whether it can improve students' performance, learning interest and learning attitude, etc., through the implementation of deep reinforcement learning teaching optimization strategy in the course of Marxism, so as to explore the optimization teaching strategy suitable for the Civic and Political Education course.

4.2. Experimental hypotheses

The purpose of this paper is to investigate the effect of subject-based teaching mode on students' learning interest, achievement and various aspects of ability, and through experimental research, the following hypotheses are proposed to be achieved:

1) The deep reinforcement learning teaching strategy based on numerical simulation can improve students' academic performance in the course of Marxism.

2) The deep reinforcement learning teaching strategy based on numerical simulation can improve students' learning interest and abilities in all aspects.

4.3. Experimental Procedures

4.3.1. Experimental design. The experiment was conducted from the end of September to the end of December 2023 with 200 students who participated in the numerical simulation teaching of the course “Marxism” at the School of Marxism of a university. The experimental teaching in the study was conducted in the school's informationized learning and management environment. The course was taught by the same instructor simultaneously to students in 4 classes divided into 2 classrooms. Both the experimental and control groups were formed by combining two classes, making no significant difference between the two groups in the dimensions related to Civics learning.

4.3.2. Experimental tools and test methods:

1) Test Papers.At the end of September 2023, a pre-test was administered to the experimental and control classes by releasing the test papers to find out the difference in the performance of the experimental and control classes before the implementation of the experiment.At the end of December 2023, a post-test was administered to the experimental and control classes by releasing the test papers

to find out the difference in the performance of the experimental and control classes after the implementation of the experiment.

2) Questionnaire. At the end of September and the end of December 2023, respectively, a questionnaire survey will be conducted on the experimental class after the implementation of teaching optimization strategies in order to find out the students' interest, attitude and motivation in learning the course of Marxism after the experiment.

3) The task statement of the subject. Released to students before the implementation of the topic, it is the goal of the implementation of the topic and the standard to test whether the topic is successfully completed.

4) Project report form. After completing the topic, students can form a general understanding of the topic process by filling out the topic report sheet, sorting out the knowledge involved in the topic activities, and recording the process and results of the implementation of the topic.

5. Effectiveness of the application of deep and intensive learning teaching strategies

5.1. Model Performance Evaluation

The experimental environment of this paper is based on the hardware device Inter Core 4.2GHz i7 7700K CPU and NVIDIA GeForce GTX 2080 Ti GPU, the software device Win10 operating system, Python 3.6 runtime environment, and the use of the open source framework Pytorch to build neural network models.

The experiments were conducted using the Junyi Academy dataset collected on the educational platform junyiacademy.org, hereinafter referred to as Junyi.The dataset contains nearly 30 million interaction records of nearly 200,000 learners on the field of Civic and Political Education and an expert-labeled Knowledge Structure Map. An interaction record contains information such as the learner number, the knowledge point corresponding to the exercise, the result of the learner's answer to the exercise, and the number of the historical interaction record group in which the interaction is located.

5.1.1. Learning Effectiveness Based on Learning Path Recommendations. In this section of the experiment for the effect of different learning path recommendation length on the learning effect, a more intuitive evaluation index - knowledge level improvement (*IKL*), i.e., the comparison of the user's knowledge level of the learning target before the beginning of the learning path recommendation and after the end of the learning path recommendation is used as a quantitative index of the learning effect, and its expression is The expression is as follows:

$$IKL = \frac{E_{end} - E_{start}}{E_{sum} - E_{start}} \tag{14}$$

where E_{start} denotes the learner's test score before the start of a learning session, E_{end} denotes the learner's test score at the end of the same learning session, and E_{sum} denotes the total score of the test.

In order to prove the validity and stability of the model designed in this paper, it is compared with the classical baseline model. The comparison models include KNN, GRU4Rec, NARM, and Q-learning.The experimental design is to compare the IKL computed in KNN, GRU4Rec, NARM, and Q-learning models when the path length N is 5, 10, 15, 20, 25, 30, 35, 40, 45, and 50, respectively.

The results of the comparison of learning effects under different models learning path recommendation are shown in Fig. 4. The experimental results show that based on the Junyi dataset, with the increasing path length, the IKL of all five models show increasing until the best, and then tend to stabilize and have a slightly decreasing trend. The difference is that for KNN, IKL has the best

performance when the path length N is taken as 25, which is 0.239. For the rest of the models, IKL has the best performance when the path length *N* is taken as 30, and according to the results of the data presented in the figure, the model of this paper has the best performance of 0.477, and IKL is higher than that of the models of GRU4Rec, NARM, Q-learning 50.47% to 101.27% higher. The above experiments prove that the IDQN model designed in this paper can achieve better recommendation results than the above four baseline models in adaptive learning path recommendation scenarios.

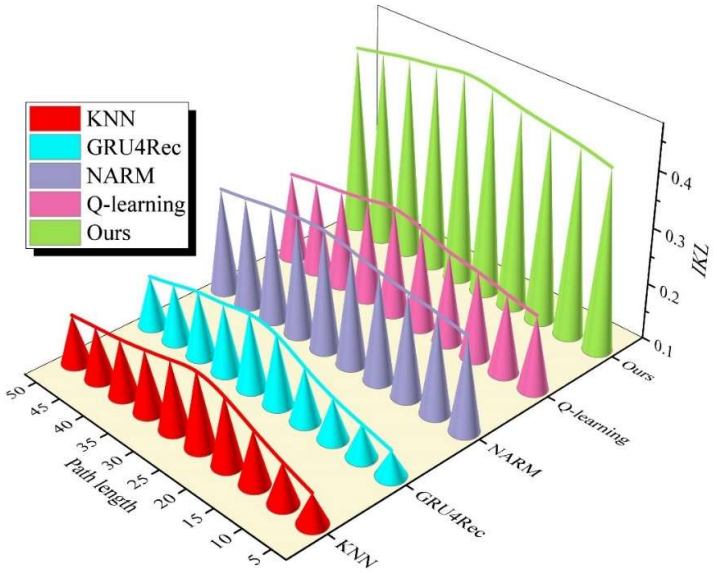


Fig. 4. The results of the study effect of different model learning paths

5.1.2. Performance of Civic Education Teaching Resources Allocation. In order to verify the effectiveness of the deep reinforcement learning-based teaching resource allocation method for civic education, the results of the allocation experiment are compared with the above method. To ensure the authenticity of the comparison experiment, the five methods are placed in the same environment, and an open Opnet platform in a scientific research institution is chosen as the operation platform for this experiment, and the performance test experiment of teaching resource allocation is carried out on the Junyi dataset.

Accuracy, recall, and F1 values are used to evaluate the performance of the teaching resource allocation methods in the experiment, and the higher the measured value, the more practical the resource allocation method is. A learning rate of 0.002 was used as the learning rate of the five allocation methods, and the continuous power allocation actions were categorized into 20 magnitudes. In order to reduce the experimental error, this experiment was conducted five times, and the final experimental results were taken as the average of the five experiments.

Figure 5 demonstrates the results of the performance comparison of different methods of teaching resource allocation. As can be seen from the figure, in the five comparison experiments, the average accuracy, recall and F1 value of this paper's method are 93.88%, 97.95% and 90.45% respectively, which are 2.57% to 21.77% higher than the comparison method. It can be seen that the optimization strategy for teaching Civics education based on deep reinforcement learning proposed in this paper can effectively improve the accuracy, recall and F1 value of the resource allocation results, and the operation process is also more stable, which can achieve efficient teaching resource allocation.

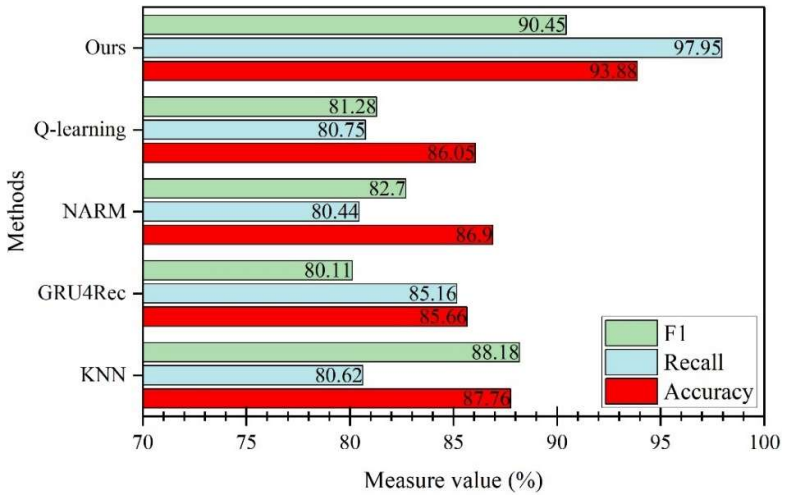


Fig. 5. The performance comparison results of the teaching resource allocation

The use of reinforcement learning algorithms may increase the chances of introducing items of interest to the user, but may also decrease the accuracy of the recommendations. For this reason, this paper uses scatterplots to evaluate accuracy and two unexpected discovery metrics. This section compares the different methods from three perspectives: accuracy, diversity, and novelty.

Figure 6 shows the accuracy, diversity, and novelty of the different methods of instructional resource allocation. The upper right corner of the scatterplot indicates the region where accuracy is well balanced with diversity and novelty. As can be seen from the scatter aggregation position in the figure, the comparison methods are more scattered, while the method of this paper is aggregated in the upper right corner of the scatter plot at the position (1,1,1). Compared with other models, the list of teaching resources generated by this paper's model is closer to the ideal region. Therefore, the list of Civics teaching resources generated by IDQN method can effectively balance the accuracy with diversity and novelty indicators.

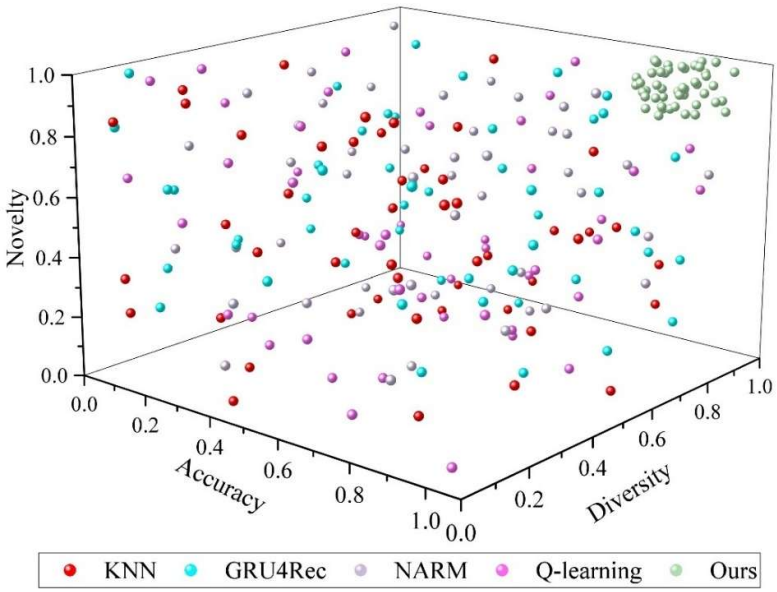


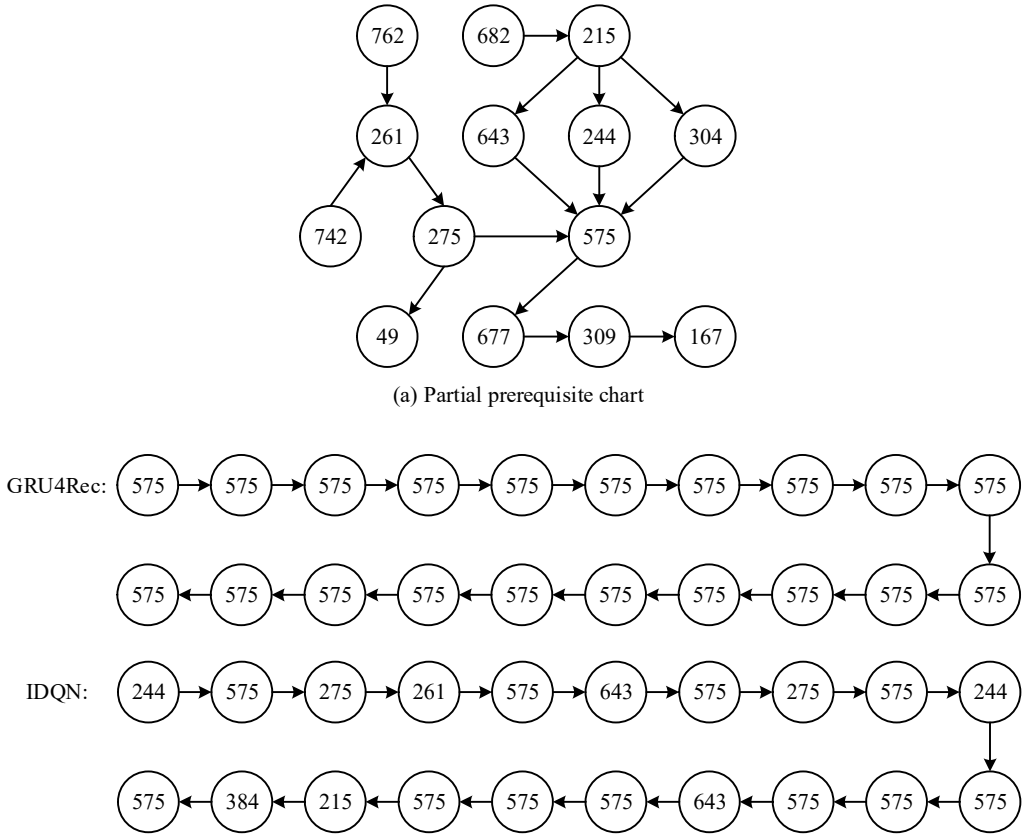
Fig. 6. The accuracy, diversity and novelty of teaching resource allocation

5.2. Evaluation of the Teaching Effectiveness of Civic Education

5.2.1. Adaptive Learning Path Recommendation. This subsection gives an adaptive learning path recommendation for the same learner using both GRU4Rec and IDQN recommendation models.

Figure 7 shows different adaptive learning path recommendations for the same learning objective. Among them, Fig. 7(a) shows a part of the prerequisite graph after preprocessing the knowledge graph contained in the Junyi dataset, and Fig. 7(b) shows the recommended learning paths for it using the two recommendation models GRU4Rec and IDQN, respectively. Among them, the learner's learning objective of the Civics course is a learning item containing knowledge point 575, and its history interaction record group is {(575,0), (575,0), (575,0), (575,0), (575,0)}, which suggests that the learner may have encountered a dilemma during the learning process, and completely failed to master knowledge point 575.

According to the experimental results, the learning path recommended by GRU4Rec for this learner contains only the learning items corresponding to the target knowledge point 575. This learner had not mastered Knowledge Point 575 originally and would not be able to complete any step in the path along this learning path. The path recommended by the model in this paper encourages the learner to review the knowledge points that have prerequisites for the target knowledge point. Based on the experimental results, it can be seen that the IDQN first recommended the learning item containing knowledge point 244 for this learner, and then chose to suggest the learner to learn the target knowledge point or the knowledge points associated with the target knowledge point according to the learner's responses, e.g., at the third time step, the IDQN recommended the learning item containing knowledge point For example, in the third time step, it recommended the learning item containing knowledge point 275, in the fourth time step, it recommended the learning item containing knowledge point 261, etc., and in the later time steps of the path recommendation, it recommended the learning item containing knowledge point 575, which shows that IDQN can recommend a more efficient and cognitively logical learning path for learners.



(a) Partial prerequisite chart

(b) Use the learning paths recommended by GRU4Rec and RL4ALPR

Fig. 7. The adaptive learning path of the same goal is recommended

5.2.2. Changes in Theoretical Achievement in Marxism. In this section, we analyze the pre-test and post-side analysis of the two groups of students' theoretical knowledge by developing theoretical test questions on Marxism, before and after the instructional numerical simulation. The difficulty of the pre-test and post-side test questions is the same, but the test questions are different, and the full score is 100 points.

Figure 8 shows the pre-test and post-side scores of the theory test of Marxism for the two groups of students. As can be seen from the figure, before the teaching numerical simulation, the distribution of the two groups of students' "Marxism" theory scores is basically the same, and most of the students' scores are in the range of 65-75 points. The average pre-test scores of the students in the control group and the experimental group are 72.84 and 72.32 respectively, and the coefficient of significance P=0.753>0.05, indicating that there is no significant difference between the average scores of the two groups of students.

After one semester of teaching the numerical simulation experiment, it can be found that the average post-side scores of the students in both groups have improved compared with the pre-test scores. The average grade of the experimental group rose to 80.26 with the intervention of learning path recommendation based on reinforcement learning, while the control group for the intervention of reinforcement learning had an average grade of 74.2, and the average grade of the experimental group was 6.06 points higher than that of the control group, with a probability of significance of $P = 0.023 < 0.05$, indicating that the theoretical grades of the experimental group and the control group were statistically significantly different from each other.

As can be seen from the achievement distribution graph, the proportion of students in the experimental group with high scores (>90) is significantly higher than that of students in the control group, and the vast majority of students' scores are concentrated in the 70-90 subsections, and the

students used have reached the 60-point passing line. In contrast, the proportion of students in the control group who did not reach the passing line was relatively high, and the proportion of students in the high score band relative to the early warning class was relatively small, in addition, the median of the experimental group's students' scores was significantly higher than the median of the control group's students' scores. Taken together, the intensive learning intervention was effective in helping students to improve their performance in the theory of Marxism.

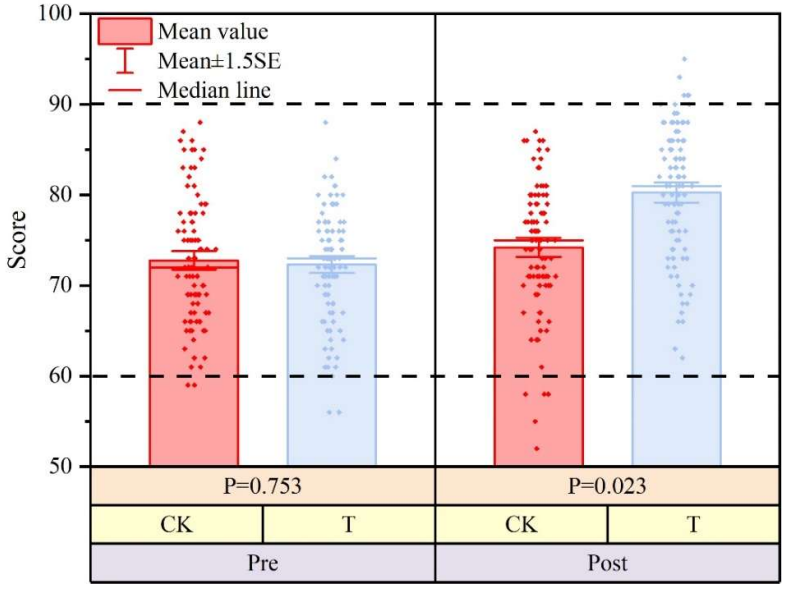


Fig. 8. Two groups of students, the former test and the back side of the test

5.2.3. Regression analysis of learning status and learning effects. It can be seen from the above that the students in the experimental group made faster progress in their theory scores in Marxism, and the theory learning effect is influenced by students' learning interest, attitude and motivation. Therefore, in this section, questionnaires on three levels of learning interest, attitude and motivation were administered to students in the experimental group before and after the experiment to explore the impact of the teaching strategy of Civics education based on enhanced learning on students' learning status. In this paper, five questions are set on each of the three levels, using a 5-level scale from 1-5, to quantify the changes in students' learning status, and the higher the score indicates the higher the students' intensity on that level.

Figure 9 shows the results of the questionnaire survey on students' learning status in the experimental group before and after the simulation. Through the questionnaire survey, it was found that the average scores of students in the experimental group in the three dimensions of learning interest, attitude and motivation before the simulation of teaching values were 3.49, 3.22 and 3.11 in that order, while after the simulation was over, the students' learning status changed significantly, and the average scores of learning interest, attitude and motivation rose to more than 4 points. It indicates that the students' learning interest, attitude, and motivation have been substantially improved under the teaching strategy of Civic Education based on reinforcement learning. This is because the teaching methods and means in the numerical simulation of teaching have been changed to improve students' interest in learning and stimulate students' initiative in learning.

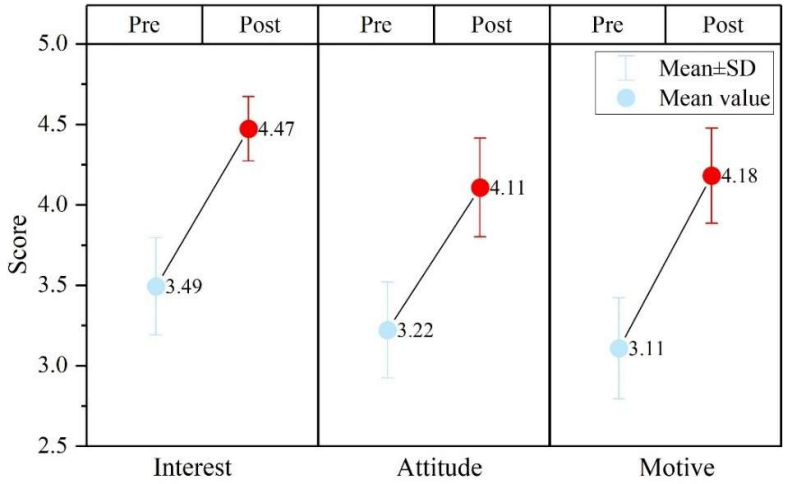


Fig. 9. Students study the results of the questionnaire survey

In this section, the regression model is used to explore the effects of learning path recommendation, teaching resource allocation, and students' learning interest, attitude, and motivation on students' learning effect of Civics course under intensive learning.

The results of the regression analysis of learning outcomes are shown in Table 1.According to the analysis of the data in the table, the linear regression analysis shows that learning path recommendation, teaching resource allocation, and students' learning interest, attitude, and motivation are significantly related to the changes in learning outcomes. The resulting regression equation demonstrates the relationship between the variables and the learning outcomes: learning outcomes = 0.027 + 0.168 x learning path recommendation + 0.217 x resource allocation + 0.178 x learning interest + 0.249 x learning attitude + 0.211 x learning motivation. In addition, the R-squared value of the model reached 0.337, which further elucidated that the five key factors of learning path recommendation, teaching resource allocation, students' interest in learning, attitude, and motivation were able to collectively explain 33.7% of the variation in the learning effect.

To ensure the stability and reliability of the model, several rigorous tests were conducted in this study. First, the model was found to be successfully validated through the F-test ($F=38.267$ and $p<0.001$), which indicated that at least one of the independent variables had a significant effect on the learning outcomes, thus confirming the reasonableness of the model. The model VIF values were tested to be below 5, indicating no covariance. The D-W value is nearly 2, showing that the data are not autocorrelated. In delving into the specific effects of the respective variables on learning outcomes, the following important conclusions were drawn: learning path recommendation, allocation of teaching resources, and students' interest, attitude, and motivation have significant positive effects on learning outcomes (regression coefficients of 0.168, 0.217, 0.178, 0.249, and 0.211, respectively, with p-values < 0.001). This means that as these four input dimensions increase, the learning effect shows a corresponding trend of improvement.

Table 1. Learning effect regression analysis results

	Nonnormalized coefficient B	Standard error	P value	VIF
constant	0.027	0.165	0.158	
Path recommendation	0.168	0.026	0.000***	1.197
Resource allocation	0.217	0.011	0.000***	2.865
Interest	0.178	0.018	0.000***	1.483
Attitude	0.249	0.039	0.000***	2.442
Motive	0.211	0.023	0.000***	2.259
Adjusted R^2	0.337			
F	38.267,P=0.000***			
D-W value	1.915			

*** stands for p significant at the 0.001 level

6. Conclusion

In this paper, the Markov decision-making process and its basic elements in reinforcement learning are analyzed in detail, and layer-by-layer optimization is carried out based on the Q learning algorithm, and a multi-intelligence deep reinforcement learning model based on the IDQN algorithm is obtained in the end. The model is applied to the numerical simulation experiment of ideology education teaching to analyze the changes in the learning effect and learning state between the experimental group and the control group, so as to evaluate the effectiveness of the model as an educational optimization strategy.

1) Model performance evaluation: The IDQN algorithm shows good learning path recommendation performance in Civics teaching, with the maximum IKL value of 0.477. And the model's accuracy, recall and F1 value for Civics education teaching resources allocation reach 90.45%~97.95%, which can realize efficient, diversified and novel teaching resources allocation.

2) Numerical simulation analysis of teaching effect: The model in this paper can can be combined with related knowledge to recommend a more efficient learning path for learners in the Civics course. After the end of teaching, the average grades of students in the experimental group and the control group show significant differences, and the experimental group is 6.06 points higher than the control group. The optimization strategy proposed in this paper improves students' interest, attitude, and motivation in Civics and Politics courses, and the learning paths and resource allocation, including all three, can improve students' learning outcomes.

References

[1] Wang, C. (2023). Exploration on the path of ideological and political education practice in colleges and universities under the new situation. Curriculum and Teaching Methodology.

[2] Vazquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement learning for demand response: a review of algorithms and modeling techniques. Applied Energy, 235(FEB.1), 1072-1089.

[3] Osakwe, I., Chen, G., Fan, Y., Rakovic, M., Singh, S., Lim, L., ... & Gašević, D. (2024). Towards prescriptive analytics of self-regulated learning strategies: a reinforcement learning approach. British Journal of Educational Technology, 55(4).

[4] Víctor Vargas-Pérez, Mesejo, P., Chica, M., & Cordón. (2023). Deep reinforcement learning in agent-based simulations for optimal media planning. Inf. Fusion, 91, 644-664.

[5] Li, X., Vasile, C. I., & Belta, C. (2017). Reinforcement learning with temporal logic rewards. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE.

[6] Catt, E., Hutter, M., & Veness, J. (2023). Reinforcement Learning withInformation-Theoretic Actuation. International Conference on Artificial General Intelligence. Springer, Cham.

[7] Mu, C., Zhao, Q., Gao, Z., & Sun, C. (2019). Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcement learning. Journal of the Franklin Institute, 356(13), 6946-6967.

[8] Ray, A., & Ray, H. (2021, May). Proposing ε-greedy Reinforcement Learning Technique to Self-Optimize Memory Controllers. In 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) (pp. 318-323). IEEE.

[9] Agostinelli, F., Hocquet, G., Singh, S., & Baldi, P. (2018). From reinforcement learning to deep reinforcement learning: an overview. University of California - Irvine, Irvine, CA 92697, USA;University of California - Irvine, Irvine, CA 92697, USA;University of California - Irvine, Irvine, CA 92697, USA;University of California - Irvine, Irvine, CA 92697, USA;.

[10] Jia, R. (2020). New Algorithms and Analysis Techniques for Reinforcement Learning. Columbia University.

[11] Xu, Z. X., Chen, X. L., Cao, L., & Li, C. X. (2017). A study of count-based exploration and bonus for reinforcement learning. IEEE.

[12] Yang, J., Wang, H., Zhao, Q., Shi, Z., Song, Z., & Fang, M. (2024, August). Efficient Reinforcement Learning via Decoupling Exploration and Utilization. In International Conference on Intelligent Computing (pp. 396-406). Singapore: Springer Nature Singapore.

[13] Alexandre, D. S. M., & Ricardo, L. D. A. D. R. (2017). An adaptive implementation of ε-greedy in reinforcement learning. Procedia Computer Science.

[14] Abuduweili, A., & Liu, C. (2022). An Optical Control Environment for Benchmarking Reinforcement Learning Algorithms. Transactions on Machine Learning Research.

[15] Ye, F., Zhang, S., Wang, P., & Chan, C. Y. (2021, July). A Survey of Deep Reinforcement Learning Algorithms for Motion Planning and Control of Autonomous Vehicles. In 2021 IEEE Intelligent Vehicles Symposium (IV) (pp. 1073-1080). IEEE.

[16] Zhu, L. (2024). A study of reinforcement learning algorithms for artistic creation guidance in advertising design in virtual reality environments. Applied Mathematics and Nonlinear Sciences, 9(1).

[17] Zhou, Y. (2022). Emergency decision-making method of unconventional emergencies in higher education based on intensive learning. Journal of environmental and public health, 2022, 4317697.

[18] Vivek S. Borkar. (2025). Some big issues with small noise limits in Markov decision processes. Systems & Control Letters105972-105972.

[19] S M Masfequier Rahman Swapno,SM Nuruzzaman Nobel,Preeti Meena,V. P. Meena,Ahmad Taher Azar,Zeeshan Haider & Mohamed Tounsi. (2024). A reinforcement learning approach for reducing traffic congestion using deep Q learning. Scientific Reports(1),30452-30452.

[20] Guangyu Yao,Nan Zhang,Zhenhua Duan & Cong Tian. (2025). Improved SARSA and DQN algorithms for reinforcement learning. Theoretical Computer Science115025-115025.

[21] Yong Gui,Zequn Zhang,Dunbing Tang,Haihua Zhu & Yi Zhang. (2024). Collaborative dynamic scheduling in a self-organizing manufacturing system using multi-agent reinforcement learning. Advanced Engineering Informatics(PA),102646-102646.