A multi-agent collaborative learning scheme for young university teachers based on reinforcement learning

Fei Jia

Nanjing Forest Police College
Nanjing, People's Republic of China

ABSTRACT: Teacher training for young university graduates involves learning about typical multi-agent learning behaviour problems within complex social system environments based on interactive forms of social reality and individual experience. Reinforcement learning is a robust method based on collaboration between multi-agents. The reinforcement learning principle and its aspects are introduced in this article, followed by a description of the Multi-Agent Markov Decision Processes (MMDP) that map young university teachers to agents. Then, the article outlines the model for reinforcement learning-based multi-agent co-operative learning. A comparison was also made between independent learning, or IL, with joint action learning, or JAL. Finally, a proposed experimental model was tested and verified. Experimental results show that the reinforcement learning model with enhanced simulation can improve the effectiveness and quality of the young university teacher's training.

INTRODUCTION

Markov decision processes, or MDPs, are named after the Russian mathematician, Andrey (Andrei) Andreyevich Markov, in which a mathematical framework is provided for modelling decision-making in situations where outcomes are partly random and partly under the control of a decision maker. Reinforcement learning was inspired by behaviourist psychology. A typical multi-agent learning behaviour problem in a complex social system environment is that of training the young university teacher. This is the dynamic process of learning and stabilising the new knowledge [1]. The process may be understood as perception - learning - decision - stabilising. The types of question posed by this process are: in this dynamic training process, what are the factors with influence? How does one initiate the training process? During learning, how do young university teachers interact with each other? What are the features that appear in the evolution of learning? Could a reverse evolution phenomenon occur? What is the incentive for it if it does appear? How would a reverse evolution phenomenon be controlled and guided? These are all issues worth exploring in depth.

To conduct in-depth research, the focus was on young university teachers and devising a new method of training that can simulate the actual situation and decision-making of these teachers. Through controlled, reproducible experiments their interactions and effects were analysed at a micro level. Changes in the guidance, motivation and management of teacher training were explored using potential models.

The concept of *an agent* in artificial intelligence, or AI, was found to greatly increase the possibility of solving problems in this type of study [2]. An agent can be viewed as a computing entity in a complex dynamic system that can autonomously perceive information in the system and act to achieve pre-set goals or tasks [3]. Similarly, in the complex system of the training of university teachers, there are those with differing characteristics that evolve through an interactive process of *perception* - *learning* - *decision* - *stabilising* in knowledge acquisition. Viewed like this, university teachers can be regarded as multi-agents with autonomous behaviour, with their own views of the world, yet able to communicate and co-operate intelligently. Therefore, the training enables the teachers, as agents, to interact with the others, thus achieving adaptive learning. This multi-agent learning system is an example of how such simulated training for teachers can simulate reality and, hence, allow the study of real problems through simulations.

REINFORCEMENT LEARNING

Reinforcement learning is an area of machine learning that was inspired by behaviourist psychology. In machine learning, the environment typically is formulated as a Markov decision process, or MDP. Multi-agent learning is an important research direction in AI [4]. Through collaboration, co-ordination and consultation, the combination of multiple agents will greatly improve the efficiency of young teachers' learning. The Internet is a multi-agent system

or MAS. An MAS is a computerised system composed of multiple interacting intelligent agents within an environment. Multi-agent systems can be used to solve problems that are difficult or impossible for an individual agent to solve. Therefore, the study of multi-agent learning appears urgent. The application of reinforcement learning in multi-agent collaboration is attracting growing interest. It combines dynamic programming and supervised learning in order to produce a powerful learning system.

As indicated above, reinforcement learning is a decision-making learning mechanism involving the interaction between an agent and a dynamic environment [5]. It is a trial-and-error type of learning method. Trial and error and delayed rewards are two most important features of reinforcement learning. The principle of reinforcement learning involves an agent that performs some type of action on the environment, changing the state of the environment and getting a reward signal from the environment to strengthen the mapping of a state and its best action. In repeating this process, an agent develops the ability to provide the best action strategy for any state of the environment [6]. This suggests reinforcement learning is highly suitable for control, planning and other strategies in the learning process. A reinforcement learning theoretical framework is shown in Figure 1.

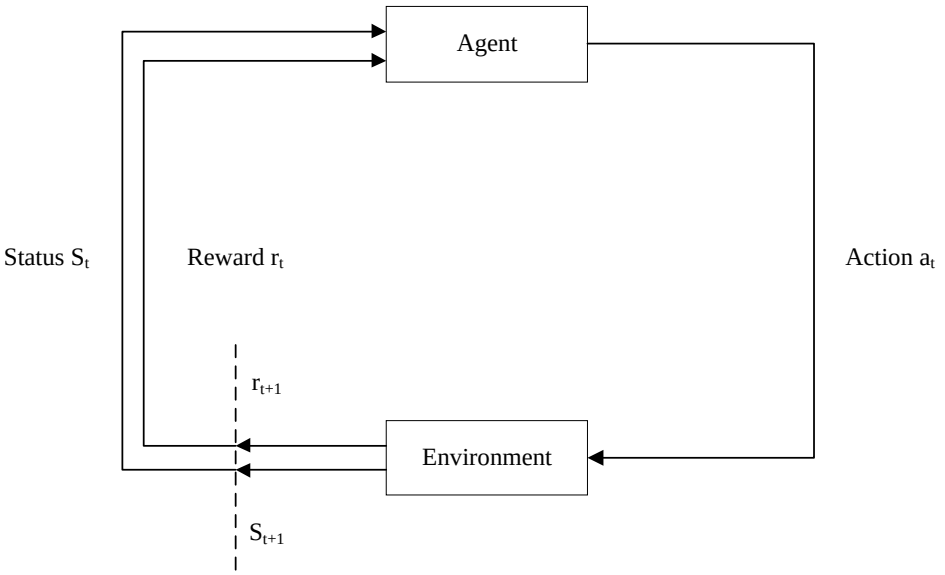

Figure 1: Schematic of reinforcement learning.

At time t, the environment status of agent is $S_t \in S$ which is the state space of the system. On this basis, agent select an action $a_t \in A(S_t)$ acting on the environment, $A(S_t)$ is the set of all possible actions the agent can take at state S_t. After the action, the environmental changes, in the next moment $t+1$, to S_{t+1}. Meanwhile, the agent receives a reward $r_{t+1} \in R$, feedback from the environment. At each moment, the agent realised from the state to the action of a selection probability mapping. This mapping is called the agent policy expressed by π_t:

$$\pi_t(s,a) = \Pr(a_t = a | s_t = s). \quad (1)$$

The agent's goal is to maximise the sum of the final rewards. At a given time, the action selected by the agent not only affects the current reward, but also the ultimate reward and the next state of the environment.

Unlike supervised learning, where the connection to learning is made mainly through the reward signal, in reinforcement learning the reward value provided by the environment is an *evaluation* of agent action, instead of telling the agent how to generate the correct action. Due to the uncertainty of the external environment, the environment itself often provides little information, and so the agent must rely on their own experience to learn. Therefore, reinforcement learning technology is considered to be an unsupervised learning technique that is more conducive to self-development in training university teachers.

UNIVERSITY TEACHER MULTI-AGENT CO-OPERATIVE LEARNING METHOD

Multi-Agent Markov Decision Processes (MMDP) and the Reinforcement Learning Algorithm

Generally, reinforcement learning problems in sequential tasks can be modelled by the Markov decision process (MDP) [7]. The MDP can be defined as a four-tuple $\langle S, A, P_r, r \rangle$, in which S and A are the set of states of the environment and the set of agent actions. The state transition function $P_r : S \times A \times S \to \Delta$, Δ is the probability distribution on the environmental state space S. If an agent performs an action a in environment s, the probability that the

environmental state changes from s to s' can be expressed as $P_r(s,a,s')$. Each time the state changes, the environment will have a reward value $r(s,a,s')$. The agent's goal is to find an optimal strategy $\pi^*(s)$ that makes the future expected discounted total $E\left[\sum_{t=0}^{\infty}\gamma R\left(s^t\middle|\pi^*\right)\right]$ revenue maximised, in which $\gamma(0\le\gamma\le1)$ is the discount rate.

Multi-agent MDP (MMDP) is an extension of the single agent MDP, which can be expressed by a quintuple $\langle\alpha,\{A_i\}_{i\in\alpha},S,P_r,r\rangle$, in which α is the set of n agent. For each $agent_{(i|i\in\alpha)}$, there is a finite set $A_{i\in n}$ of actions. Joint actions $\langle a_1,a_2,\cdots,a_n\rangle$ taken by the agents make up the joint action space $A=\times A_i$. The definition of S and $P_r r$ are essentially similar to the MDP, the only difference is A corresponds to the combined actions in MMDP.

The *Q*-learning algorithm [8] has been mainly used in this article in reinforcement learning. Delimit $Q^*(s,a)$ as the discount reward when the agent takes action a in state s. Agent's optimal strategy $\pi^*(s)$ is the action that maximises the value of Q in state s. The definition of Q^* and π^* is consistent with the Bellman principle of optimality, which is that any optimal strategy is necessarily constituted by the optimal sub-strategy. The solution to Q^* and π^* is as follows:

$$Q^*(s,a)=r(s,a)+\gamma\sum_{s'\in S}P_r(s,a,s')\max_{a'}(s',a') \tag{2}$$

$$\pi^*(s)=\arg\max\left(Q_w^*(s,a)\right) \tag{3}$$

Reward function $r(s,a)$ and State transition function $P_r(s,a,s')$ are unknowns. *Q*-learning uses the value function iteration algorithm to approximate the optimal solution, that is:

$$Q_{t+1}(s,a)=(1-\beta)Q_t(s,a)+\beta\left(r(s,a,s')+\gamma Q_t(s')\right) \tag{4}$$

The theory proved that when the learning rate β satisfies certain conditions, the *Q*-learning algorithm will be able to converge to the optimal solution.

Multi-agent Collaborative Learning

In multi-agent collaborative learning, the agent using the *Q*-learning algorithm obtains reward values by interacting with the environment and the *Q* value is updated according to the reward value. Several main modules in the agent reinforcement learning model are shown in Figure 2.

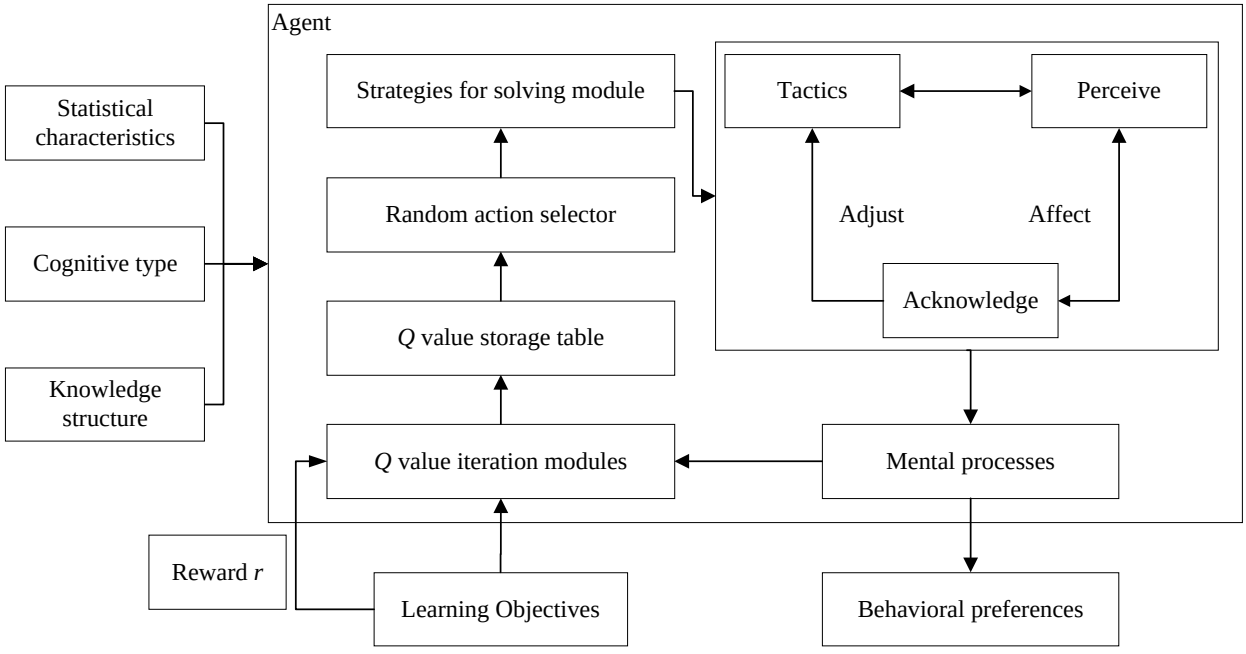

Figure 2: The framework of the multi-agent co-operative learning method.

The framework can be explained as follows:

- *Statistical characteristics:* in studies using mathematical statistics, the fact that young university teachers are concerned about the overall composition of each individual is not in itself significant, but associated with this is the requirement to investigate the relevant characteristics of each individual in the overall distribution between situations.
- *Cognitive type:* different cognitive styles of young teachers mean their learning will have different characteristics. The user's own field of independent reinforcement learning characteristics is significantly higher than the other field-dependent users. Therefore, agent settings should be adjusted based on the cognitive learning outcomes.
- *Knowledge structure:* a teacher's agent should have not only depth of expertise, but also a broad range of optimised knowledge reflecting career development.
- *Q value iteration modules:* according to the environmental status *s*, update *Q* value by agent's action *a* and reward value by instantaneous feedback.
- *Q value storage table:* store *Q* value and update the *Q* value storage table after each iteration of the *Q* value.
- *Strategies for solving module:* according to the current state of the input s and the estimation of the *Q* value in the state, solve for the optimal strategy.
- *Random action selector:* according to $\in$*-greedy* strategy selection mechanism, except using 1-ε for selecting the optimal strategy, it can also select other non-optimal strategies with probability ε.
- *Strategy, results, cognitive:* a teacher's agent will do a determining and adjustment of mental models based on the feedback and interaction between agents.

This model assumes the agent chooses strategy j in round t; the probability $p(j,t+1)$ of acceptance strategies *j* in round $t+1$ will be:

$$p(j,t+1) = p(j,t) + \alpha^{BM} \cdot (1 - p(j,t)) \qquad j = d(t) \wedge \pi(t) \geq 0 \tag{5}$$

$$p(j,t+1) = p(j,t) - \beta^{BM} \cdot p(j,t) \qquad j = d(t) \wedge \pi(t) < 0 \tag{6}$$

When the agent does not choose strategy k in round t, the probability $p(k,t+1)$ of acceptance strategies k in round $t+1$ will be:

$$p(k,t+1) = p(k,t) - \alpha^{BM} \cdot p(k,t) \qquad k \neq d(t) \wedge \pi(t) \geq 0 \tag{7}$$

$$p(k,t+1) = p(k,t) + \beta^{BM} \cdot (1 - p(k,t)) \qquad k \neq d(t) \wedge \pi(t) < 0 \tag{8}$$

$d(t)$ refers to the agent's actual selection strategy, $\pi(t)$ means optimal strategy to pay, α means the reinforcement strength of the correction output (i.e. $\pi(t) \geq 0$), β refers to the reinforcement strength of negative output (i.e. $\pi(t) \leq 0$), α and β are constants, with a value between 0 and 1.

SIMULATION AND RESULT ANALYSIS

Design of Experiments

Young university teachers were mapped to agents, producing a multi-agent reinforcement learning model. Observing the differences in learning behaviour, mental characteristics and learning rules characteristic of young university teachers, the agents produced background knowledge and a cognitive background through the intervention of the Reinforcement Learning Method. In addition, grouping the agents by cognitive style and further observing the different cognitive styles revealed learning differences.

Analysis of Results

Table 1 has the data from the young university teachers' multi-agent behaviours, from which were extracted on each *round*, the agents' proportion of preferred *individual learning*, the *collaborative learning* ratio, and by the reinforcement learning stimulus from *single individual learning* into the *collaborative learning* ratio.

From each round of data, figures were gathered for *enter collaborative learning*; *collaborative learning stimulated by the reinforcement learning ratio* and *collaborative learning stimulated directly by reinforcement learning*.

Table 1: Young university teacher multi-agent behaviour data.

Learning methods	Proportion and number of rounds
Individual learning ratio	82.4%
Collaborative learning ratio	34.7%
Ratio of collaborative learning stimulated by reinforcement learning	52.9%
Rounds of enter collaborative learning	4.44
Rounds of collaborative learning stimulated by reinforcement learning ratio	5.74
Rounds of collaborative learning stimulated directly by reinforcement learning	1.3

As is shown in Table 1, the university teachers' agents clearly preferred *self-learning* to *collaborative learning*. But after several rounds of reinforcement learning, this changed. Through self-exploration of trial and error, the young university teacher multi-agent, eventually, will learn better through the *collaborative learning* method, and will make great progress. These results are explained by the reinforcement learning model, which can improve the effectiveness and quality of young teachers' training.

CONCLUSIONS

In this article, a collaborative learning model for young university teachers was proposed, based on reinforcement learning. In the model outlined above, the multi-agents co-ordinate their action selection through reinforcement learning combined with a co-ordination mechanism, so as to achieve better collaborative learning results. In this article a comparison was also made of independent learning, or IL, with joint action learning, or JAL, in multi-agent systems based on reinforcement learning collaboration. On this basis, the new collaborative learning model was summarised, and experiments conducted, which verified the effectiveness of the model.

ACKNOWLEDGEMENTS

This work was supported by *Fundamental Research Funds for the Central Universities* (No. RWYB201339).

REFERENCES

1. Ko, A.H.R., Sabourin, R. and Gagnon, F., Performance of distributed multi-agent multi-state reinforcement spectrum management using different exploration schemes. *Expert Systems with Applications*, 40, **10**, 4115-4126 (2013)
2. Cubillos, C., Diaz, R., Urra, E., Cabrera-Paniagua, D., Cabrera, G. and Lefranc, G., An agent-based solution for the Berth allocation problem. *Inter. J. of Computers Communications and Control*, 8, **3**, 384-394 (2013).
3. Pang, S.N., Ban, T., Kadobayashi, Y. and Kasabov, N.K., LDA merging and splitting with applications to multiagent cooperative learning and system alteration. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetic*s, 42, **2**, 552-564 (2012).
4. Daneshfar, F. and Bevrani, H., Load-frequency control: a GA-based multi-agent reinforcement learning. *IET Generations Transmission and Distribution*, 4, **1**, 13-26 (2010).
5. Skatova, A., Chan, P.A. and Daw, N.D., Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Frontiers in Human Neuroscience*, 7, ID: 525 (2013).
6. Ozogul, G., Johnson, A.M., Atkinson R.K. and Reisslein, M., Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Computers & Educ.*, 67, 36-50 (2013).
7. Mannor, S. and Tsitsiklis, J.N., Algorithmic aspects of mean-variance optimization in Markov decision processes. *European J. of Operational Research*, 231, **3**, 645-653 (2013).
8. Banerjee, B. and Kraemer, L., Action discovery for single and multi-agent reinforcement learning. *Advances in Complex Systems*, 14, **2**, 279-305 (2011).