2D-02

Playing *mini-Hanabi* card game with Q-learning

Fei Tong [1,3] Masahiro Ichiki[2,3] Kenichi Nakazato[3]
[1]Graduate school of Informatics, Kyoto university
[2]Graduate school of Informatics, Nagoya university
[3]Bosch Center for Artificial Intelligence, Corporate Research, Bosch Corporation

Abstract--Hanabi card game is a cooperative card game. Unlike the other games, the players cannot see their own cards and can only see other people's. So, it is very challenging for AI players to learn this game. In this study we simulated the Hanabi card game and trained the AI player by using the Q-learning method. However, Q-learning method will take a large amount of time if space states is numerous. Therefore, we parameterized the numbers and kinds of cards to estimate the size of the space states. Finally, we minimized the cards parameters and trained the AI player by using Q-learning in a short time.

I. Introduction

Since the birth of Artificial Intelligence (AI), it has been closely tied with games. That is because people generally think that the process of humans playing games contains human intelligence. Therefore, when people create a program that can complete some kind of human games, we think that this program also has some kind of "intelligence". So many games such as chess and poker game often serve as challenge problems, benchmarks, and milestones for the progress of AI.

Past successes in such benchmarks, for example chess [1], have been limited to two-player. However, in recent years, researches have been extended to multiplayer [2] which is a recognized AI milestone. Then Hanabi card game is predicted to be another AI milestone, because, unlike the other games, players cannot see their own cards and can only see other people's. This means the AI agents have to learn the cooperation with each other.

The Hanabi deck contains cards in 5 colors (white, yellow, green, blue, and red). The values on the cards to be dealt are 1, 1, 1, 2, 2, 3, 3, 4, 4, 5 for each color. On a player's turn, you must complete one, and only one, of the following 3 actions: I. Give a hint, II. Discard a card, III. Play a card.

To give a hint, the player has to take a blue token which have 8 blue tokens in total. The player can then tell a teammate information about only one color or only one value that the teammate has in his hand. This action cannot be performed if there are no blue tokens.

Discarding a card, a blue token will be returned. This action cannot be performed if there are already 8 blue tokens.

To play a card, it is successful if the card is a 1 in a color that has not yet been played, or if it is the next number sequentially in a color that has been played. If the player plays a wrong card, he will receive a red token. The game is over if players receive 3 red tokens.

II. Methods

We first simulated the Hanabi card game. And then trained the Hanabi card game by Q-learning [3] which is a model free reinforcement learning algorithm. The goal of Q-learning algorithm is to estimate action values under certain states. In the Hanabi card game there are only 3 actions -- I. Give a hint, II. Discard a card, III. Play a card, so it is easy to simulate. However, there are too many states because there have 50 cards in total. It means, if every player has n cards in hand, there will have ${}_{50}C_n$ states for each player. However, numerous space states mean large Q-table which will take a large amount of time during training. To solve the problem, we minimized and parameterized the colors and values of the cards.

We assumed there are N colors. Each card has value, up to L, on it. Multiple numbers of the same type cards are permitted, e.g. red-1,1,1,2,2,3. The total number of cards satisfies with the equation, $T \geqq LN$, and there are M players, each player has n cards in hand. So, the hand cards' state of each player can be represented by S_j .

Here, our state space S_j only shows player has a card j or not. That is, we ignore the number of a specific card j in this description, for simplicity. Therefore, the total number of states is $2^{LM.}$ Then the Q-value can be updated by,

$$Q^{new}(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha\left(R(s,a) + \gamma \max_a Q(s_{t+1}, a)\right). \quad (2)$$

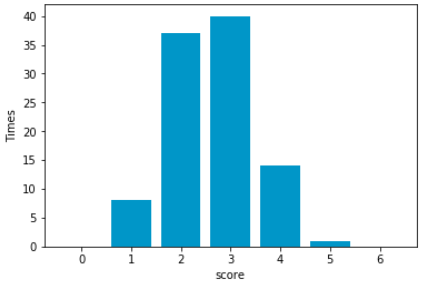
R(s, a) is the reward received when moving from the state s_t to s_{t+1}, and α is the learning rate. γ is the discount factor which determines the importance of future rewards. s is the state and a is the action. In our experiment we set learning rate α as 0.1 and discount factor γ as 0.9. If a correct card is played we give a high reward otherwise the reward is 0.

We minimized the colors and values of the cards, on the other hand we changed the rules a little to increase the difficulty of the game. In the action, III. Play a card, can be played if the card is a 1 in a color that has not yet been played, or if it is the next number sequentially in a color that has been played. We changed the action, III. Play a card, rules to card can be played if the card is a 1 in a color that has not yet been played, or if it is the next number sequentially in a color that has been played by the previous player. We name it mini-Hanabi card game.

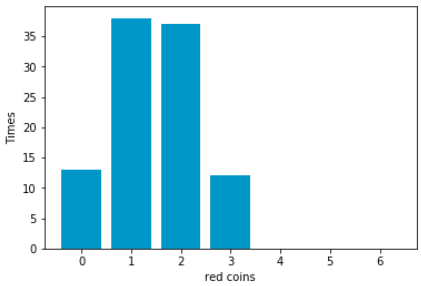
III. Results

In our simulation we supposed colors N=2, the values on the cards of each color are 1, 1, 1, 2, 2, 3. The total number of cards is T=12 and players number M=2, each player has n=3 cards in hand.

We used the score to evaluate the Q-learning method. In order to calculate their scores, we add up the largest value card for each of the 2 colors, so the total points of mini-Hanabi card game is 3 points(red) + 3 points(blue) = 6 points. We played the game 100 times, the scores and red coins distribution is shown as Fig.1. The mean score of 100 times is 2.63 point which means almost played 3 cards correctly. The mean value of the obtained red coins is 1.48 which means play 1~2 wrong card in a game.



(a)



(b)

Fig.1 Distribution of (a)scores and (b)red coins

IV. Conclusions

We minimized the colors and values of the cards, on the other hand we increased the difficulty of the game and named it mini-Hanabi card game. In this paper we used Q-learning method to train AI agent play this game. The mean score of our proposed method is 2.63 points, and the total score of this game is 6 points. In future work, the Deep Q-learning Network (DQN) method is needed for training the numerous colors and values of *Hanabi* card game.

References

[1] D. Silver, J. Schrittwieser, K. Simonyan, A. Huang, I. Antonoglou, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. Nature, 550:354– 359, 2017.

[2] N. Brown and T. Sandholm. Superhuman AI for multiplayer poker. Science, 11, 2019.

[3] C. J. C. H. Watkins. Learning from delayed rewards. PhD thesis, University of Cambridge England, 1989