A REINFORCEMENT LEARNING MODEL FOR PLAYING A FISHING CARD GAME "TABLIĆ"

Nikola N. Kilibarda, Mladen M. Ravlić, Vladimir M. Milovanović

Department of Electrical Engineering, Faculty of Engineering,
University of Kragujevac, 6 Sestre Janjić Street, 34000 Kragujevac
e-mail: vlada@kg.ac.rs

Abstract

Reinforcement learning saw a tremendous rise in use over the past decade particularly in robotic and general control applications, as well as in board and video games where a series of remarkable results have been achieved. This fundamental machine learning paradigm is especially suitable for problems in which the environment can be at least partially modelled and/or simulated faster than in real time. One such setting can be seen when playing card games. This paper proposes a reinforcement learning methodology for training an intelligent agent that plays Tablić, a popular fishing-style card game. The bots trained within various finite Markov decision process scenarios characterized by different parameters have been compared both between each other and with a plain greedy algorithm. It is demonstrated that properly constrained reinforcement learning agents which favour long-term over short-term rewards can explore and inherit non-trivial moves. Results prove that even simple Q-learning algorithms surpass greedy agents in expected win rate and can reach and presumably exceed human-level performance. With minor modifications, other (fishing) card games can utilize the same open library which was developed over the course of this work.

Keywords: Reinforcement learning, card game, Tablić, Q-learning, greedy algorithm.

1 Introduction

Learning through interaction with the environment is one of the first forms of learning a person adopts and uses during the course of their life. Even without a dedicated teacher, humans have a direct sensory motor connection to their environment. Using this connection they can deduce much about the cause and consequences of their actions as well as what needs to be done to achieve their goals. During a person's life, such interactions are undoubtedly their primary source of knowledge in regards to their surroundings, as well as to themselves. Regardless of whether a person is learning to ride a bicycle, play the piano or hold a conversation, they are pretty much aware of their environment and how it responds to their actions, and through their behaviour they try to influence it.

Reinforcement learning [1] represents one of the three basic paradigms of machine learning that is based on interaction with the environment. Recently its popularity peaked, especially in video game bots [2, 3] that can even outperform humans. In this paper the principles of machine learning are applied to the popular playing card game *Tablić* with the goal of creating a model that learns to play the game by playing against itself. Even though reinforcement learning has been applied to various other card games previously [4], to the author's knowledge, there are no published works where reinforcement learning was implemented on the Tablić game.

2 Learning Tablić

2.1 The game of Tablić

Tablić is played with a standard deck of 52 cards without the jokers. Each round, both players draw 6 cards and take turns playing them in a sequence by placing them on the table. If the value of the played card is equal to the value of a card on the table, or the sum of multiple cards on the table, they may claim the played card and any number of cards with the same value and/or sets of cards and place them in their pile. The King (K) has a value of 14, the Queen (Q) 13, and Jack (J) 12. The Ace (A) can have the value of both and 11, and it simply depends on the player's choice.

The players' scores are determined as follows [5]:

- The cards 10, K, Q, J, A are worth 1 point.
- The 2 of clubs is worth 1 point.
- The 10 of diamonds is worth 2 points.
- The player who ends the game with the most captured cards earns 3 points.
- Players earn one point each time they clear the entire table.

2.2 Q-learning

Reinforcement learning is an area of machine learning where learning is done through the interaction with the environment. It requires no prior knowledge of the rules that govern the environment, nor any marked input/output pairs.

A reinforcement learning model's behaviour can be represented by its *policy* $\pi(a|s)$, a probability distribution that specifies the probability of the model taking action a while in the state s. The goal of reinforcement learning is for the model to find an *optimal policy*, which is a policy that maximizes the reward function by assigning a probability of 1 to the optimal action in any given state, and a probability of 0 to all other actions. In the case of Tablić, the reward function is the number of points the player has at the end of a game.

The model decides its next move by examining all actions available from the current state and choosing the one that is expected to yield the highest long-term reward. The expected reward gained from taking action a in state s is called the *Q-value* $Q(a|s)$. Thus, the goal can be defined as learning to accurately estimate the Q-value of any given action in any given state, which is done by using the *Q-learning* [6] algorithm. The algorithm is as follows:

Initialize $Q(a|s)$ arbitrarily;

Repeat (for each episode):

Initialize state s;

Repeat (for each step of episode):

Choose action a from state s using policy derived from Q;

Take action a, observe reward r and new state s';

Update Q-values for visited states s and actions a according to (1);

$s \leftarrow s'$;

Until s is the terminal state.

The equation used to update Q-values is:

$$Q(a,s) \leftarrow (1-\alpha) \cdot Q(a,s) + \alpha[r + \gamma \cdot \max_{a'} Q(a',s')] \quad . \tag{1}$$

The hyperparameter α in the above equation is the *learning rate* which determines to what extent new information overrides old one, while γ is the *discount factor*, which serves to devalue future rewards to allow for convergence of the value function in environments with infinitely long episodes.

Q-learning is an off-policy reinforcement learning algorithm, as the policy being learned and used for value evaluation is different from the one used to select the next action. A partly-random Q-based policy is used for exploration, most commonly the ε-*greedy* policy, which has a probability of ε to select a random action in order to explore more of the state-space, a probability of 1-ε to choose the action estimated to be the optimal one.

3 Implementation

For the purposes of model evaluation, a library for Python was developed and is now publicly [7] available. Two types of players are available as part of this library.

3.1 Greedy player

The greedy player implements greedy Tablić strategy. This strategy looks at all the available moves and chooses the one that immediately scores the most points while not looking at the long-term benefits. In cases where two or more moves score the same amount of points, it will choose a move according to the strategy that scores the most amount of cards.

3.2 The Reinforcement Learning player

The Reinforcement Learning player is a player who is trained using the Q-learning algorithm. Function $Q(a,s)$ is modeled using a feed forward neural network that consists of an input layer of a size 80, two hidden layers of a size 160 with the ReLU activation function and the output layer of the size of 1. The input for the neural network represents the observation vector of the game (cards in hand, cards on the table, cards taken by both players, player that's on the move, and which player collected the last cards).

This player is using the ε-greedy strategy with ε being 1 when the training starts and this value slowly decreases as the player gains more knowledge. During testing the ε is equal to 0, so the player always plays the best moves.

4 Results

For the purpose of this analysis, seven different models were trained and compared with the greedy player and amongst themselves. All the players were trained using the same parameters, the same structure of the neural network, and the same number of episodes (50000). The only difference between the players is the discount factor γ.

Table 1 shows different players compared with the greedy player. Looking at the data we can see that all the trained players score more points on average than the greedy player even though they are also very greedy. Also it can be seen that players trained with a higher value of the discount factor γ perform better against the greedy player

Player	wins	draws	losses	player points	opponent points	points greediness	points & cards greediness
Greedy Player	0	1000	0	26.214	26.214	100%	100%
Reinforced Player $\gamma = 0.00$	535	32	433	26.828	25.341	99.8432%	95.2796%
Reinforced Player $\gamma = 0.50$	592	31	377	27.912	25.359	99.8294%	91.8676%
Reinforced Player $\gamma = 0.75$	620	33	347	28.617	24.703	99.7583%	95.3605%
Reinforced Player $\gamma = 0.85$	685	39	276	29.257	24.158	99.6546%	96.8058%
Reinforced Player $\gamma = 0.90$	665	32	303	29.141	24.124	99.5742%	97.6538%
Reinforced Player $\gamma = 0.95$	748	32	220	29.894	23.347	99.1764%	98.4632%
Reinforced Player $\gamma = 1.00$	732	38	230	30.170	23.232	98.6907%	97.6720%

Table 1: Results of different Reinforcement Learning players against the plain Greedy Player.

and simultaneously exhibit lower greediness, and this can be interpreted and viewed as a sign of intelligence. However, the described reinforcement learning algorithm behavior is a direct derivation of the model feature that values long- instead of short-term rewards.

Next, the players are compared amongst themselves, and the results are shown in Figure 1. The same conclusion can be drawn here, too. The players trained using a higher discount factor score more points against the players who have a lower discount factor.

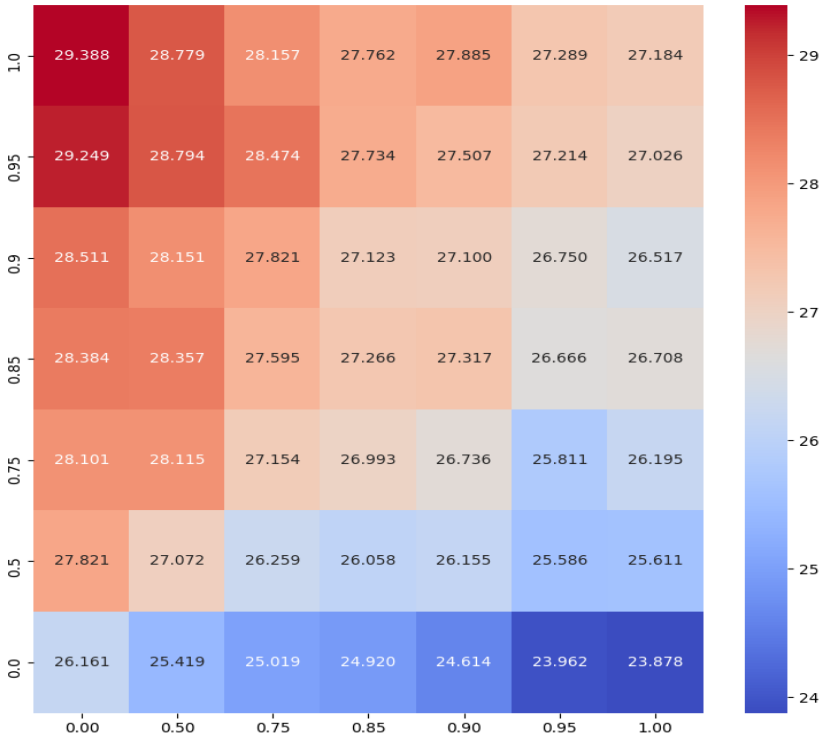

Figure 1: Performance of the RL players trained with different discount factor against one another.

5 Conclusion

Using a simple reinforcement learning algorithm, such as Q-learning, it is possible to create a model that surpasses a greedy algorithm when playing a game of Tablić. This showcases the positive aspects of reinforcement learning when trying to maximize long-term instead of short-term results. Most importantly this was done without teaching the model or giving it any clue or prior knowledge on how to play the game. It was simply given the rules of the environment (game) and simulated interaction which shows how the game is played. This means that all what is needed to train a good player using reinforcement learning is a good model of the environment (in this particular case Tablić), and the model will learn everything by itself.

The generality of the Q-learning algorithm also makes it easy to use this project to create players for other card games, and it is possible to achieve this without major modifications to the training algorithm.

References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, New Jersey: Pearson, 2020.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *NIPS Deep Learning Workshop*, 2013.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[4] D. Zha, K.-H. Lai, Y. Cao, S. Huang, R. Wei, J. Guo, and X. Hu, "RLCard: A toolkit for reinforcement learning in card games." in *AAAI-20 Workshop on Reinforcement Learning in Games*, Feb. 2020.

[5] J. McLeod, "Rules of card games: Tablić," www.pagat.com/fishing/tablic.html, accessed: April 15, 2022.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, Massachusetts: MIT Press, 2018.

[7] N. Kilibarda, M. Ravlić, and V. Milovanović, "A reinforcement learning model for playing Tablić card game," www.github.com/milovanovic/TablicRL, accessed: April 15, 2022.