争奪型の報酬のある三人対戦の不完全情報ゲームにおける強化学習

Reinforcement Learning Agents Playing a Three-Player Imperfect Information Board Game with Robbable Rewards

平田　旺樹
Ouki Hirata
法政大学情報科学部コンピュータ科学科
E-mail: ouki.hirata.4f@cis.k.hosei.ac.jp

Abstract

Board games are used as a test ground for reinforcement learning AI agents because of their expressiveness of real issues, abstract goals, and complex rules. Reinforcement learning is weak at some board games with imperfect information and delayed rewards, e.g., Ticket to Ride (T2R). Although previous work solved some of the weaknesses, it treated only 1-vs-1 situations. In T2R played by more than two agents, it is important how much resources consume rewards that other agents can rob. Rewards of this type are difficult for reinforcement learning because it can only get rewards after its action, not in the enemies' turns. This paper treats such rewards by making reinforcement learning agents get variable rewards when agents get robbable rewards. More concretely, a reward is changed by the rate of progression of a game, the percentage meaning whether an agent can hold a robbable reward at the end of the game, or both. In addition, this paper modifies the custom map that the previous work adapted. It also verifies win rates without robbable rewards, and resource losses of the agents that get simple rewards and the agents that get variable rewards in round robin tournaments in the modified map.

1. はじめに

ボードゲームは実世界での問題を切り抜いて表現・反映できる．例えばボードゲーム CATAN には他者と双方に利益があるように交易をしながら自分の利益を最大にする行為が含まれる．そのため，ボードゲームは AI の研究の試験場としても扱われている．その中で，不完全情報ゲームは様々な状況が想定でき，従来の学習方法では難しい題材もあるため，深く研究されている．実際に，報酬が遅延して獲得する種類のボードゲームの先行研究がある [1]．しかし，強化学習の多くの研究では二人対戦のゲームを扱っており，三人以上のゲームを対象とした研究は少ない [2] [3] [4]．三人以上のゲームの研究が難しい理由は，状態数が増えるだけではなく，獲得人数に上限があり，ゲーム中にその報酬の保持者が変動する争奪型の報酬があることが挙げられる．

本研究ではボードゲーム Ticket to Ride (T2R)を題材とし，三人でのプレイを想定した環境で強化学習で勝率を高めることを目的とする．T2R で三人以上の対戦時には争奪型の得点への関与の仕方が戦略に関わる，そのため，争奪型報酬によって，報酬を得た際に学習エージェントに与える報酬を可変にする手法を提案する．具体的には，経過時間，現状の争奪報酬獲得可能性により変動させる．これにより，単純の報酬の与え方を行った AI より優れた勝率を記録するかを検証する．それに付随して，先行研究で用いられていた縮小マップに争奪型報酬を適切に組み込む．

2. 関連研究

本研究では Yang ら研究 [1] (以下，先行研究)をもとにして進めていく．先行研究は，遅延した報酬が得られるゲームとして T2R を扱い，この課題を普遍的な学習方法で解決している．具体的には，ルールベース AI に頼らない自己学習の強化学習で他の学習エージェントに高い勝率を記録した．対戦環境は二人対戦であり，独自に考案した縮小マップを使用した．先行研究の今後の課題として挙げられたのは，状態数が多いオリジナル版のマップでの学習，三人以上の対戦環境での学習，ニューラルネットワークの使用の三つである．

3. 準備

3.1. Ticket to Ride

Ticket to ride は二人から遊べる不完全情報ゲームのボードゲームである．盤面には都市とそれらを繋ぐ線路が存在し，自分の目的の都市を繋ぐ線路を保持することを目的とする．各プレイヤーには列車カードと行先チケットの 2 種類の手札があり，それらは非公開情報である．列車カードは線路を保持するために使用し，行先チケットに記された二つの都市を繋ぐと得点になる．行先チケットは山札から，列車カードは山札か 5 枚の表向きのカード(公開情報)から補充できる．

各プレイヤーの得点は線路を保持した際に長さに応じて加算される．またゲーム終了時に，行先チケットの内容を達成できた場合それに記された数だけ得点が，達成

Supervisor: Prof. Hiroshi Hosobe

できなかった場合は同量の失点がある．また，ゲーム中に一筆書きで最長の道を繋いだプレイヤーには追加の得点が加算される．最長の道を繋いだプレイヤーが更新された場合，その得点は新たなプレイヤーへ移動する．

3.2. 強化学習

強化学習とは，状態s_tに対する行動a_tに評価値$Q(s_t, a_t)$や，行動の出力ルールである方策$\pi(s_t, a_t)$があり，方策$\pi(s_t, a_t)$を更新し，最適な方策πを模索する学習方法である．1 回の行動a_t毎に報酬rを受け取り，方策πを更新することで最適な方策を学習する．報酬は行動をした瞬間にしか与えることができないため，それ以外のタイミングで得点が変動するゲームは強化学習では難しいとされる．

4. 提案手法

強化学習の仕様上，報酬rを与えるタイミングを変更することはできない．そのため，与える報酬rを状況に応じて可変にする手法を提案する．また，プレイヤー人数の変化に伴う学習環境の縮小化も行う．

4.1. 報酬値の可変化

学習中に争奪型の報酬を得た際に，ゲームの進行状況と争奪報酬の保持確率に応じて与える報酬を増減させる手法を提案する．強化学習の仕様上，争奪報酬を失った際に負の報酬を与えることができないため，与える報酬をその時の価値で増減させる必要があると考えたからである．以下の 3 種類の方法で争奪報酬獲得時の報酬を可変にする．なお，実験では提案手法をそれぞれ別のモデルで学習させ，比較を行う．

4.1.1. 経過ターンに応じた可変化

ゲーム序盤は戦局を見極めることが難しいため，争奪報酬が与える影響を小さくする必要がある．現在ターン数t，前試合の総ターン数t_p，争奪報酬基礎点r_bを用いて，与える報酬rを次式のように表す．

$$r = r_b \frac{t}{t_p}$$

ただし，t_pの初期値を 1 とする．試合の進行度を計る場合，試合にかかる平均ターン数を用いた方が望ましい．しかし，それはエージェント単体で測れる情報ではないため，前回試合のターン数で代用する形をとった．

4.1.2. 保持確率に応じた可変化

ゲーム終了時に争奪報酬を保持できる可能性が低いにも関わらず，争奪報酬にリソースを割くことは賢い戦略とは言えない．そのため，保持確率によって報酬を変化させることは効果があると考えられる．保持確率P_k，争奪報酬基礎点r_bを用いて，与える報酬rを次式のように表す．

$$r = r_b P_k$$

争奪報酬保持確率P_kが小さい場合，争奪報酬より他の報酬を優先する動きをすると考えられる．争奪報酬保持確率P_kが大きい場合，争奪報酬を他の報酬より優先する動きをすると考えられる．そのため，自分が争奪報酬を得られる確率に応じて争奪報酬へのリソースの割き方が変わることが予想される．

争奪報酬保持確率P_kは強化学習を用いて導出する．保持確率学習エージェントは試合学習エージェントを用いて学習を行う．試合の盤面をランダムに生成し，その時点での保持確率を予想する．その後，試合学習エージェントで試合を進行し，結果と確率の乖離度に応じて報酬を与える．試合学習エージェントが一定回数学習を行った後に保持確率学習エージェントを学習させる．これを交互に繰り返し学習を進めていく．

4.1.3. 経過ターンと保持確率に応じた可変化

上記二つの手法を複合した式を提案する．現在ターン数t，前試合の総ターン数t_p，保持確率P_k，争奪報酬基礎点r_bを用いて，与える報酬rに関する以下 2 つの式を提案し，どちらが優れているか実験を行う．

$$r = \frac{1}{2} r_b \left(P_k + \frac{t}{t_p} \right), \qquad r = \frac{r_b t P_k}{t_p}$$

前者は平均をとり，後者は乗算をする形である．

4.2. 縮小マップの拡張

先行研究では，状態数と行動数を大幅に削減するために縮小マップが開発された．これにあわせて，即時報酬や遅延報酬の得点も調整を行った．本研究でも縮小マップを用いる．しかし，縮小マップでは争奪報酬がマップスケールに合わせた変更がされていなかった．そのため，争奪報酬の得点を変更する必要がある．

4.2.1. マップ分析

先行研究では，様々なバージョンの T2R を用いて分析を行い，妥当であるといえる縮小マップの作製を行った．ただし，本研究では著者の事情により Ticket To Ride–America (T2RA)の一バージョンのみで分析を行った．この分析を基に争奪報酬の得点を設定する必要があった．表 1 は T2RA と縮小マップにおける争奪報酬の得点と他の得点源，およびそれらに関連する要素との比較である．

オリジナル版の争奪報酬の分析として，オリジナル版の争奪報酬は行先チケットの中央値と近い値をとっている．争奪報酬の得点は行先チケットの最高得点より低く，最低得点より高い．オリジナル版と縮小マップの比較として，各要素は 0.22 倍から等倍の範囲に縮小している．倍率の中央値は 0.36 であり，平均は 0.45 である．

表 1　オリジナル版と縮小マップの分析と比較

	T2RA	縮小マップ	倍率
争奪報酬	10		
最大線路長	6	3	0.50
最小線路長	1	1	1.00
列車トークン数	45	10	0.22
行先チケット最高得点	22	5	0.23
行先チケット最低得点	4	2	0.50
行先チケット平均得点	11.43	3.75	0.33
行先チケット中央値	10.50	4	0.36

4.2.2. 報酬設定

本研究では争奪報酬のゲーム内での得点であるr_bを 4 点とする．ゲームの体裁を保つうえで得点は整数であることが望ましい．そのため，縮小マップでの争奪報酬の得点は 3 点か 4 点が妥当であると考えた．前述の通り，オリジナル版の争奪報酬は行先チケットの中央値と近い値をとっている．よって 4 点がふさわしいと判断し，本実験ではその得点で実験を行った．

5. 実装

言語は Python を用いる．使用する主なライブラリは StableBaseline3 と gymnasium である．実装は学習環境，強化学習アルゴリズムの改良，可視化を行った．強化学習アルゴリズムは Proximal Policy Optimization (PPO) [5]と呼ばれる学習方法を用いる．提案手法以外の箇所の実装は先行研究と同様のものになるように努めた．学習環境は gymnasium の Env クラスを継承した子クラスを作成した．強化学習アルゴリズムはライブラリ StableBaseline3 のクラス PPO，ActorCriticPolicy，Categorical を継承した子クラスを作成した．報酬以外の方策等の箇所は既存のプログラムを利用した．行動マスキング処理に必要なメソッドをオーバーライドし，処理を加えた．また，PPO には保持確率予測エージェントを学習させる処理，自己学習を行わせる処理を追加した．

6. 実験

6.1. 実験方法

試合学習エージェントの学習中の学習時間，得点取得数の変遷と保持確率予測エージェントの学習中の獲得報酬を記録する．先行研究の学習回数と同様に 1000000 回の学習を行う．また，保持確率予測エージェントの学習スパンを変更したモデルで比較を行い，適切な切り替えスパンを検討する．切り替えスパンは，1000 回，10000 回，100000 回で行う．

争奪報酬で与える報酬を変更した強化学習モデルで総当たりを行い，勝率，勝利時の争奪報酬獲得割合を記録する．勝利時の争奪報酬獲得割合は，争奪報酬に対する戦略の変化に対する判断材料とする．モデルは提案手法の四つに加え，争奪報酬が不変のモデルを二つ用意する．それらは争奪報酬の得点だけ報酬を与えるモデルと一切の報酬を与えないモデルである．以下では，一切の報酬を与えないモデルを MinAI，争奪報酬の得点だけ報酬を与えるモデルを MaxAI，4.1.1 節の方法を採用したモデルを TimeAI，4.1.2 節の方法を採用したモデルを PredictAI，4.1.3 節の方法を採用したモデルをそれぞれ ComplexAverageAI，ComplexMultiplicationAI とする．

6.2. 実験結果

6.2.1. PredictAI の学習スパンの設定

表 2 は学習スパンを変えた PredictAI の学習時間，表 3 は提案手法外の AI との勝率である．また，図 1 は保持確率予測エージェントの，図 2 は試合学習エージェントの学習過程である．学習スパンが長い程試合学習エージェントの学習効率が良く，勝率が高い．また，1000 回スパンのエージェントは学習過程が安定しておらず，常に学習を続けている．これより後の実験で保持確率予測エージェントを用いた AI は全て学習スパンを 100000 回に設定している．

表 2　PredictAI の学習時間

切り替えスパン	1000 回	10000 回	100000 回
時間(s)	81149	19102	9533

表 3　各 PredictAI と MinAI，MaxAI との勝率

	MinAI/MinAI	MinAI/MaxAI	MaxAI/MaxAI
/1000 回	0.4869	0.36155	0.3533
/10000 回	0.4911	0.35724	0.3572
/100000 回	0.5304	0.37917	0.3713

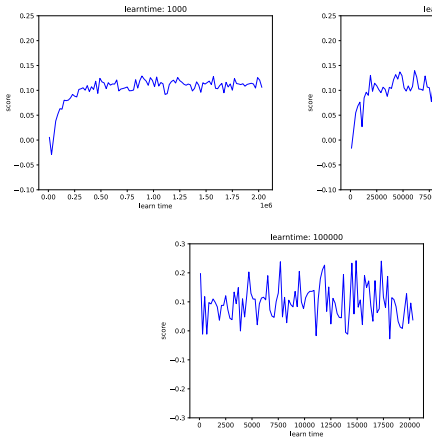

図 1　各 PredictAI の保持確率予測エージェントの学習過程

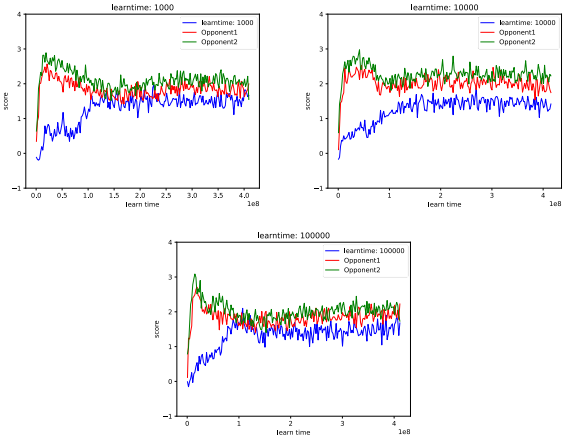

図 2　各 PredictAI の試合学習エージェントの学習過程

6.2.2. 学習データ

表 4 は各試合学習エージェントが 1000000 回学習する際にかかった時間である．学習にかかる追加の時間はお

よそ 30 分ほどに収まっている．図 3 は各試合学習エージェントの学習中の得点推移である．MaxAI が最も学習の効率が高いが，得点が他より低い傾向がある．次点で学習効率が良いのは PredictAI である．また，ComplexMultiplicationAI の方が ComplexAverageAI よりも学習効率が良く，平均的に獲得得点も多い．

表 4　学習時間の計測結果

	Min-AI	Max-AI	Time-AI	Predict-AI	Complex-Average-AI	Complex-MultiplicationAI
時間	7805	8034	7878	9533	10102	9864

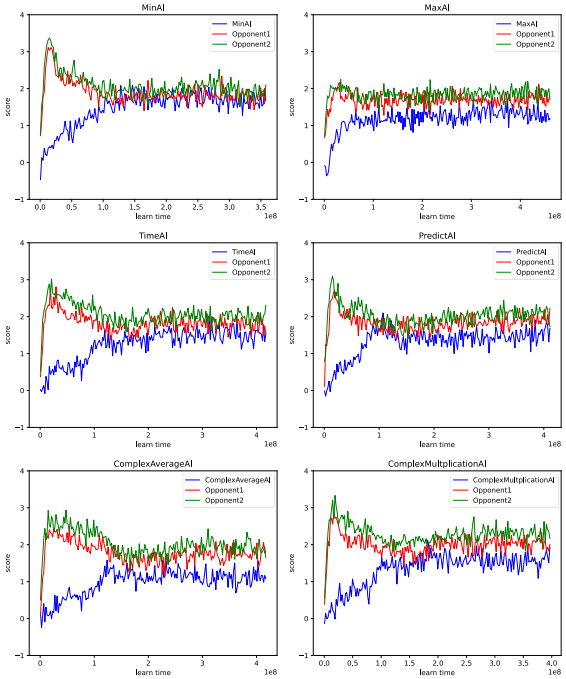

図 3　各エージェントの学習過程

6.2.3. 総当たり戦

図 4 は各エージェントで総当たり戦を行った際の勝率である．横軸の AI の勝率を表す．TimeAI，PredictAI (図中では PredAI)，ComplexAverageAI (図中では CAAI)は MinAI，MaxAI より高い勝率を修めている．PredictAI が最も高い勝率を記録している．ComplexMultiplicationAI (図中では CMAI)は MaxAI より勝率が低い．

図 5 は勝利した試合の内，争奪報酬を獲得せずに勝利した試合の割合である．ComplexMultiplicationAI が最も争奪報酬を獲得せずに勝利している．他の提案手法の AI も MaxAI よりは争奪報酬がなくとも勝利している．また，ComplexAverageAI は TimeAI，PredictAI より争奪報酬を獲得せずに勝利している．

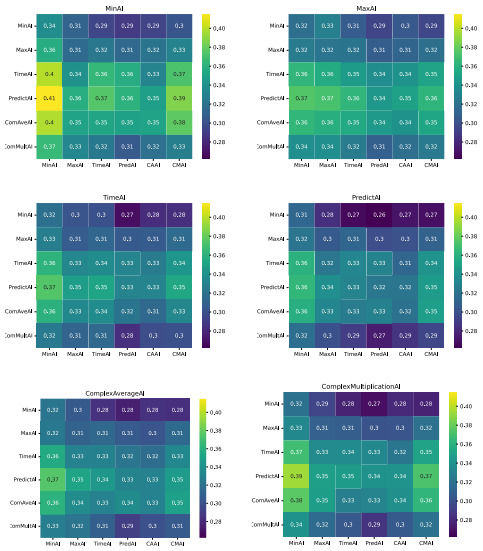

図 4　1v1v1 の総当たり戦の勝率

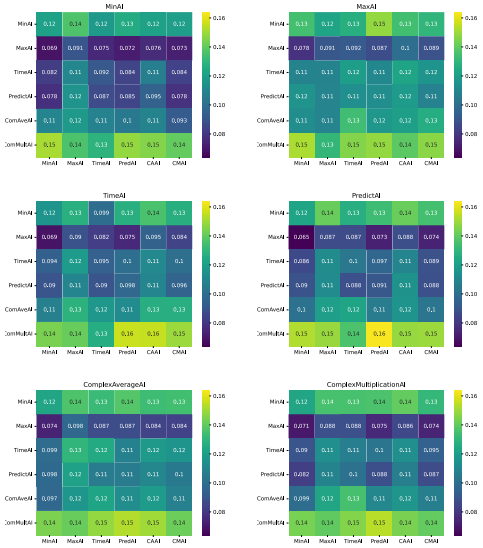

図 5　勝ち試合の内争奪報酬未獲得の割合

7. 議論

6.2.1 節で学習スパンが長い程結果が良い理由として，学習序盤の十分な学習時間の確保が考えられる．学習序盤は動きがランダムに近く，その状態で評価値が変動してしまうことが学習の妨げになっていると考えられる．ただ，今回の実験では 100000 回より長いスパンで検証を行っていない．そのため，学習スパンが長い程結果が良いと決定づけるにはさらなる検証が必要である．

6.2.2 節で保持確率予測エージェントを用いた AI の学習効率について，保持確率予測エージェントの学習前はそれがノイズになってしまっている可能性が考えられる．学習前でも保持確率予測エージェントは報酬の設定に関与している．そのため，ほとんどランダムな値が報酬設

定に影響を与えているため，学習を妨害していると考えられる．

6.2.3 節で 4.1.3 節の AI がどちらも PredictAI より勝率が低かったことについて，単純な式で提案手法を融合したことが考えられる．全体的に PredictAI のほうが TimeAI より勝率が高いため，中間をとる式だと経過ターンに応じた可変化が足を引っ張っている可能性が考えられる．ただ，TimeAI も勝率が低い訳ではないため，TimeAI と PredctAI を比較し，互いの優れた部分だけを抽出できる式をたてれば勝率を高くできると考えられる．

ComplexMultiplicationAI は争奪報酬に頼らない勝率が一番高かった．これは，争奪報酬に関する設定を行っていない MinAI より勝率が高い．そのため，争奪報酬を避けながら勝利するという戦略がとれていると考えられる．よって，報酬の可変化による争奪報酬に対する戦略の変化は行えているといえる．また，ComplexAverageAI も他の提案手法よりは争奪報酬に頼らない勝率が一番高い．そのため，複数のパラメータによって報酬を変化させることは効果があると言える．よって，争奪報酬に対する報酬の可変化に使用するパラメータを変更・調整すれば，勝率が高く，かつ争奪報酬に寄らずとも勝利できる AI を作成できると考えられる．

8. おわりに

本研究では，争奪報酬獲得時にエージェントに与える報酬を可変にした．勝率の向上，争奪報酬を獲得せずに勝利する割合の向上は見られたが，両立することは適わなかった．

今後の課題として以下が挙げられる．

- 縮小マップにおける争奪報酬のより緻密な検討
- 等間隔以外の学習スパンの検討
- ルールベース AI，人間を用いた強さの提案手法の評価
- 三人対戦とそれ以上の人数の学習が同様のものであるかの検討

これらを検討することでより高い勝率の AI を作成できると考えている．また，実際の試合を用いて分析を行うことで，可変時に必要なパラメータを再検討できると考えられる．

文　献

[1] S. Yang, M. Barlow, T. Townsend, X. Liu, D. Samarasinghe, E. Lakshika, G. Moy, T. Lynar and B. Turnbull, "Reinforcement Learning Agents Playing Ticket to Ride—A Complex Imperfect Information Board," *IEEE Access*, vol. 11, pp. 60737-60757, 2023.

[2] C. Huchler, "An MCTS agent for ticket to ride," Maastricht Univ, M.S. thesis, Fac. Humanit Sci., The Netherlands, 2015.

[3] F. D. M. Silva, S. Lee, J. Togelius and A. Nealen, "AI as Evaluator: Search Driven Playtesting of Modern Board Games," *Proc. AAAI Workshops*, pp. 959-966, 2017.

[4] V. Lind and L. Strömberg, "Board game AI using reinforcement learning," Örebro Univ, B.S.thess, Sch. Sci. Technol., Örebro, Sweden, 2022.

[5] J. Sculman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv*, no. 1707.06347, 2017.