1L2-4in1

多人数戦略すごろくゲームにおけるマルチエージェント学習による協調的動作の実現手法の考察

Multiagent Reinforcement Learning Methods for Realizing Cooperative Behavior in Multi-Player Strategic Railway Management Simulation Game

杉浦生隼 *1　福田直樹 *2
Kihaya Sugiura　Naoki Fukuta

*1静岡大学総合科学技術研究科
Graduate School of Integrated Science and Technology

In this paper, we discuss the methods of reinforcement learning to realize agents who behave cooperatively to other players in multiplayer strategic railway management simulation game. We define the game as multiagent environment and we created the leaning agents who learned by the seven evaluation functions to cooperatively play with each other in railway management simulation game. We conduct the experiments to examine the result of playing the game that includes imitated different level user player agents.

1. はじめに

近年ではさまざまな種類のコンピュータゲームが登場し，複数人でプレイするゲーム環境において，他のプレイヤーを積極的に攻撃しなければ勝つことが難しいという特徴を持つゲームが登場した．これらのゲームは友人同士でのプレイにおいて，過剰に他のプレイヤーを攻撃することによってプレイヤーを不快にしてしまいプレイヤーが快適にプレイできないことがある．

本研究では，このような特徴を持ったゲームのルールを踏まえたうえで，マルチエージェントの技術を用いてプレイヤーを代替するエージェントがユーザプレイヤーに協調し，複数のレベルのユーザが快適にプレイできるような学習プレイヤーエージェントの組み合わせの検討を行う．

ゲームに参加している他のプレイヤーの手持ち金との差を参照するエージェントをゲームに参加させることで，擬似ユーザプレイヤーを適度に拮抗したゲーム状態に置くことができるかどうかを検討した．

2. 背景

2.1 有限 n 人繰り返し囚人のジレンマ問題

ボードゲームはルール上複数のプレイヤーでプレイされることが多く，各プレイヤーの行動選択の結果，各プレイヤーの利益が相反するような状態となる事がある．

二人のプレイヤーが協調または非協調を選択する単純な囚人のジレンマモデルでは，非協調の支配戦略が存在することが知られている．この支配戦略は相手の意思決定によらない．すなわち，相手の意思決定戦略を知らずに自分の意思決定を行うことができるということである．

これに対して n 人繰り返し囚人のジレンマモデルは，複数のプレイヤーによる囚人のジレンマモデルであり，単に一度だけの協調非協調のモデルではなく複数のプレイヤーの意思決定が複数回にわたって行われる進化的なモデルである [鈴木 01]．

n 人繰り返し囚人のジレンマは現実社会のモデルに近い．現実世界において，意思決定を無数に繰り返すことは少なく，時間などの制限によって繰り返し回数は有限となることが多い．同様にゲーム環境においても，終了条件が設定されていることが多く，プレイヤーの行動選択には時間的な制限がある．

連絡先: 杉浦生隼，静岡大学総合科学技術研究科，〒 432-8011 静岡県浜松市中区城北 3-5-1, ia12049@s.inf.shizuoka.ac.jp

2.2 本研究における戦略的すごろくゲーム

本研究における戦略的すごろくゲームは，マルチプレイヤーのボードゲームである [杉浦 16]．図 1 に，本研究で扱う，戦略的すごろくゲームの模式図を示す．本ゲームはコンピュータ上でプレイされる．複数のプレイヤーがターンごとに複数の駅からなる路線図を模した環境をさいころの目に従って動きまわる．また，プレイヤーは黄色の駅に止まることにより，さいころをふる代わりに他のプレイヤーにデメリットを与える攻撃的行動をとることを選択できる．

本研究で扱う多人数戦略すごろくゲームでは，プレイヤーの利益となる行動が必ずしも他のプレイヤーの不利益にならない．全プレイヤーが同じ手持ち金で同順位でゲームが終了することが起こり得る．このことからプレイヤー同士は協調して，同順位でゲーム終了にならないように，他のプレイヤーを攻撃することができる．

一般的なボードゲームにおいて，プレイヤーの意思決定は自分の番ごとに行われることが多く，囚人のジレンマモデルのようにプレイヤーの意思決定が同時に行われることは少ない．このゲームは非ゼロ和ゲームである．

プレイヤー プレイヤーは自分のターンごとにサイコロをふるか，手持ちのカードの権利を行使して他のプレイヤーを攻撃できるするかを選択できる．プレイヤーの目的は自身の手持ち金を最大化することである．

駅 駅には,「青の駅」,「赤の駅」，および「黄の駅」の 3 種類がある．青の駅はプレイヤーの手持ち金を増やす．赤の駅はプレイヤーの手持ち金を減らす．黄色の駅はプレイヤーに攻撃的な行動を行う権利を与える．

攻撃的行動 黄色の駅に止まって権利を得ていた場合，自分のターンの開始時にさいころをふることに加えて，攻撃的行動を行うかを選択できる．攻撃的行動を行うと他のプレイヤーにデメリットを与える行動を取ることができる．

2.3 マルチエージェント学習手法の適用

環境に複数の学習エージェントが存在する環境において，学習エージェントが協調的動作を獲得するための学習手法は大きく 2 通りのものに分けられる．

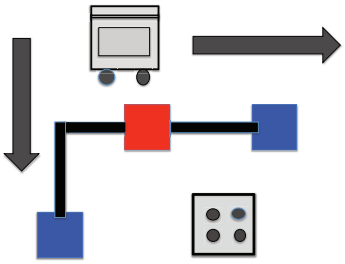

図 1: 戦略的すごろくゲームの模式図

一つは，同じ目的を持つ学習エージェント同士が互いに通信をし，状態を参照しながら学習エージェント同士がその目的を達成するために都合の良い行動を学習する手法である．具体的にはゲーム理論を用いて学習を行う手法が提案されている [Hu 15]．この手法では，学習エージェントが環境上において達成すべき目的を決め，学習エージェントの位置や状態を確認し合いながら協調的動作を実現する．

もう一つは，必ずしも目的が同じであると明示しない学習エージェント同士が，学習エージェントの設計者が用意した評価関数をもとにして，「自然に」協調性を獲得する手法である [Wicke 15]．

3. 多人数戦略すごろくエージェント評価プラットフォームの実装

本研究を行うための実験環境として，実装に Node.js と WebSocket プロトコルを用いた．ゲームオブジェクトの保持に Node.js 上で実行可能な WebSocket サーバを用意し，同様に Node.js 上で動作するプレイヤーエージェント (WebSocket クライアント) を実装する．一つのゲームサーバに対して複数の (人間またはエージェント) プレイヤーが接続を行う．環境に依存しにくい，マルチプラットフォームでの動作を考える．実際にブラウザ上でゲームをプレイ可能なアプリケーションを，図 2 と図 3 に示すように，作成した..

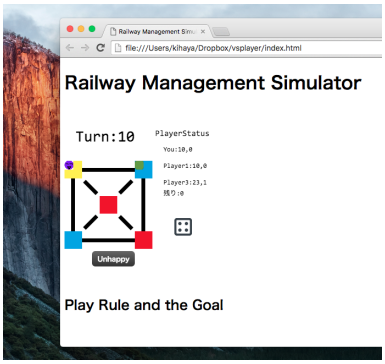

図 2: 戦略的すごろくゲームの実行例

4. 協調的動作獲得のための評価関数の設計

ここでは実験的なゲーム環境上における Q-Learning[Sutton 98] エージェントの行動獲得に着目する．表 1 にゲームの条件設定を示す．ゲーム環境における順位の変動を実現するために，プレイヤーを 3 体とする．1 体は Q-Learning により学習を行うエージェントであり，2 体はランダムな行動選択を行うエージェントとする．これは学習エージェントが存在し，他のプレイヤーが存在するゲーム環境としての最小の構成である．

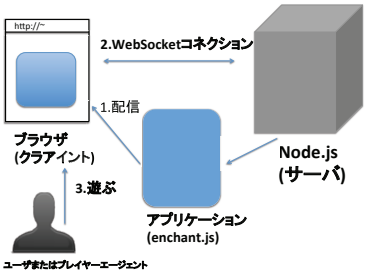

図 3: 実験ゲーム環境のアーキテクチャ

ゲーム環境には 5 つの駅からなるマップを用意する．プレイヤーが青駅に止まった場合プレイヤーの手持ち金を +1 し，プレイヤーが赤駅に止まった場合プレイヤーの手持ち金を −1 する．

プレイヤーが攻撃的な行動をとることを選択した場合，その対象となったプレイヤーの手持ち金を −2 する．学習エージェントは自分のターンが来ると，さいころをふるか可能であれば攻撃的な行動を行うかを ϵ-Greedy 法 (ϵ=0.1) によって選択する．さいころをふることを選択した場合，さいころの出目に応じて最終的に止まることが可能な駅を列挙しそれらを選択肢として行動選択を行う．攻撃的な行動を行うことを選択した場合，どのプレイヤーに対して攻撃的な行動をとるのかを選択する．1 ゲーム=99 ターンとし，7000 ゲーム行った．

表 1: 実験ゲーム環境

青色駅の数	赤色駅の数	黄色駅の数	プレイヤー数
2	2	1	3

Q-Learning エージェント (学習エージェント) が学習を行うために学習エージェントの状態をそのゲームにおける現在の順位とする．

学習エージェントは学習の過程において，ゲーム環境から報酬を得る．この実験において 3 種類の異なる報酬の与え方を設定する．これらの報酬は同時に学習エージェントに与えるのではなく，個々に異なる評価関数としてエージェントに与える．以下に報酬の詳細を示す．

この実験において複数の異なる報酬 (評価関数) の与え方を設定する．以下に報酬の詳細を示す．

評価関数 A 学習エージェントがとった行動に応じた報酬を与える．青駅，黄駅に移動する行動か攻撃的行動を選択した場合に正の報酬 (いずれも同一の値) を与える．赤駅に移動する行動を選択した場合，負の報酬を与える．

評価関数 B 学習エージェントの行動後の順位に応じた報酬を与える．1 位を最大の報酬とし，3 位が最小となるように与える．

評価関数 C 学習エージェントが行動後に，他プレイヤーの手持ち金を参照して，どちらか他のプレイヤーの手持ち金

が自分の手持ち金と等しいなら正の報酬を与える．それ以外の行動には負の報酬を与える．

評価関数 D 学習エージェントが黄駅に止まった次の自分のターンに攻撃的行動を選択した場合に負の報酬を与える．それ以外の行動選択には正の報酬を与える．

評価関数 E 学習エージェントが行動後に，他のプレイヤーの手持ち金を参照して，どちらかの手持ち金と自分の手持ち金の差が 2 以内ならば正の報酬，それ以外の行動には負の報酬を与える．

評価関数 F 学習エージェントの順位が変動した場合に報酬を与える．順位が上がった場合に正の報酬を，下がった場合に負の報酬を与える．

評価関数 G 学習エージェントが攻撃的行動を選択した場合に負の報酬を与え，攻撃的行動を選択しなかった場合に正の報酬を与える．

表 2 より，評価関数として順位の変動値を報酬とした場合では，攻撃的な行動を選択した回数と黄駅に止まった回数の平均の合計が赤駅に止まった割合を超えていることがわかる．また，学習エージェント自身が不利になる行動，すなわち赤駅に止まる行動価値は割合が減るが，現在の順位を報酬とした時よりも小さくはならないことが観測された． 一方で，現在の順位に基づいた報酬を与えた場合には，学習エージェントが黄駅に止まる割合が順位変動を報酬とした場合よりも多いためその割合が順位変動をもとにしたものより高いが，常に攻撃的行動をとる選択とはなっていない．

この攻撃的行動の選択の割合は，順位を考慮せずに攻撃的行動の有無のみを報酬とした場合とは異なることが観測された．

表 2: 実験ゲーム環境における行動選択回数の評価関数ごとの割合 2

評価関数	青駅	赤駅	黄駅	攻撃的行動	平均順位
評価関数 A	48.2%	7.25%	43.8%	0.69%	1.1
評価関数 B	49.3%	3.9%	24.0%	22.6%	1.0
評価関数 C	50.5%	43.0%	6.1%	0.4%	1.9
評価関数 D	24.8%	46.5%	27.9 %	0.6%	2.6
評価関数 E	39.9%	42.1%	14.0 %	3.8%	1.9
評価関数 F	44.84%	22.60%	20.89%	11.67%	1.29
評価関数 G	25.68%	47.11%	26.56%	0.65%	2.68

ゲーム環境上において学習エージェントが行動価値を獲得できるかを検討するためのシミュレーション環境を構築し．学習エージェントの報酬を決定する評価関数が，他のプレイヤーに対する攻撃的な行動の選択に与える影響について述べた．予備実験として，強化学習時の報酬を決定する評価関数を用意し，評価関数による行動選択の学習のされかたの違いについて考察した．順位の変動値を報酬とした場合で，学習エージェントが自身の順位をむやみに下げてしまわまないような行動が学習により得られていることを観測した．現状の条件設定において攻撃的な行動の選択が報酬の与え方により異なることが観測できた．

より複雑な環境下で同様の振る舞いを強化学習によって必ずしも獲得できるかという点と，その振る舞いが，人間のプレイヤーから見て思いやりのある協調行動であるとみなされるかどうかという点についての検討は，今後の課題である．

5. 学習エージェントの協調性の評価

5.1 実験条件

ここでは，どのような学習エージェントがゲーム環境上でプレイしている場合にユーザプレイヤーに対して協調性を獲得できるのかを確認する．前述の学習エージェントの行動獲得の実験において，評価関数 A と評価関数 B の学習エージェントは学習の結果，平均順位が 1 位に近い値となった．このエージェントの行動選択手法を工夫することで，現実のユーザプレイヤーのレベルを模倣することを考える．

擬似プレイヤーは，プレイヤーの順位に応じた報酬を与える評価関数によって学習した Q 値に従いながら行動を行うエージェントである．順位に応じた評価関数によって学習したプレイヤーエージェントは高い順位を維持できるような行動を学習したエージェントであるといえるが，実際のゲーム環境でユーザプレイヤーがゲームをプレイする場合，様々な要因から，必ずしも最適な行動選択を行わないことが予想される．これを再現するために，擬似プレイヤーに実際のゲーム環境でプレイするユーザプレイヤーの行動選択を模倣するように，確率 ϵ でランダムな行動選択を行うように設定する．

順位に応じた報酬によりゲーム環境を学習したエージェントが行動選択に用いる ϵ の値を変更することにより，ユーザプレイヤーのような振る舞いをするエージェントを模倣する．ユーザプレイヤーのような振る舞いをするエージェントは，確率 ϵ で自身が必ずしも得をしない行動を選択してしまうような行動選択をするエージェントである．

この学習エージェントとユーザプレイヤーが対戦することを想定する場合，このゲームに慣れていないユーザプレイヤーはゲームが難しく感じられ不快感を感じる可能性がある．

学習したプレイヤーエージェントの組み合わせを検討し，複数のエージェントを用いたゲームプレイの実験を行った．ユーザを模倣したプレイヤーエージェントとして，順位に応じた報酬により学習したエージェントを用いて，行動選択の ϵ の値を 0.1，0.3，0.5 と設定した．ユーザを模倣したプレイヤーエージェントの対戦相手として，前述の実験において異なる評価関数を用いて作成した学習エージェントを用いた．1 ゲーム=99 ターンとして (各プレイヤーの行動選択で 1 ターン)，500 ゲーム行った．

手持ち金の差によって学習を行ったエージェントをプレイさせることによって，他の他のプレイヤーに対して協調的な行動を取ることができるのではないかと考えた．ここでいう協調的行動とは各プレイヤー間の手持ち金の差が開かないようにする行動のことである．

5.2 結果

図 4 にユーザプレイヤーを模したプレイヤーエージェント，報酬の組み合わせにより学習したプレイヤーエージェントと差が開いた際に負の報酬を与えたプレイヤーとを対戦させた際の各プレイヤーの手持ち金の推移を示す．戦略的なユーザプレイヤーを模倣した ϵ=0.1 のプレイヤーエージェント A が報酬の組み合わせによって学習した戦略的なプレイヤーエージェント B および手持ち金の差が開いた際に罰を受けることにより学習したプレイヤーエージェント C と対戦した場合，約 7 ターン目で戦略的なユーザプレイヤーを模したプレイヤーエージェント A の手持ち金が 3 プレイヤーの間で最大となり，ゲーム終了時まで最大を保ち続けた．

図 5 と図 6 に示すように，ユーザプレイヤーの行動選択の不確実性が大きくなったケースでは，プレイヤーエージェント B が 20 ターン前後まで，手持ち金の差は開かなかった．

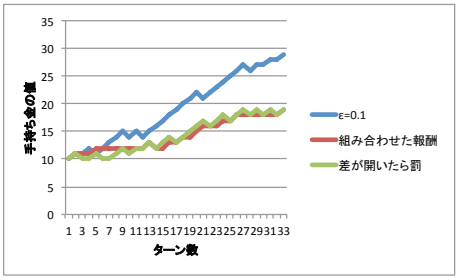

図 4: 順位に応じた報酬 (ϵ=0.1)，報酬の組み合わせ，差が開いた際の報酬にもとづいたプレイヤーエージェントの各プレイヤーの手持ち金の推移 (1 ゲーム)

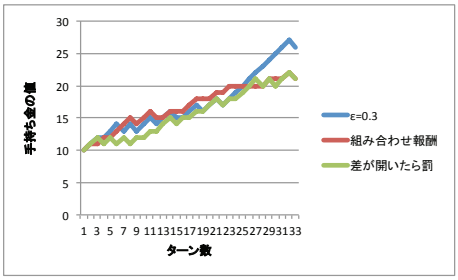

図 5: 順位に応じた報酬 (ϵ=0.3)，報酬の組み合わせ，差が開いた際の報酬にもとづいたプレイヤーエージェントの各プレイヤーの手持ち金の推移 (1 ゲーム)

5.3 考察

ϵ=0.1 の強いユーザプレイヤーを擬似プレイヤーと設定した場合には，序盤のうちから 3 プレイヤーの中で最大の手持ち金を保っており，他のプレイヤーを圧倒するような結果が得られた．現実のユーザプレイヤーが，報酬の組み合わせと手持ち金の差が開いた際の報酬によって学習したエージェントと対戦する場合，ゲーム環境においてプレイヤーの順位の変動が発生しづらいことからゲームを退屈に感じる可能性があるが，ユーザプレイヤー以外の 2 人のプレイヤーの手持ち金は拮抗しており，順位変動が起きやすい状況にある．

ϵ=0.3 の中間的なユーザプレイヤーを擬似プレイヤーと設定した場合には，終盤まで 3 プレイヤー間の拮抗状態が続いており，25 ターン前後で，擬似プレイヤーの手持ち金が 3 プレイヤー間で最大となっている．差が開いた際の報酬によって学習したエージェントが中間的な実力のプレイヤーと対戦する場合に，手持ち金の差をうめるような行動を選択するため手持ち金が拮抗している可能性がある．

プレイヤーエージェント C はどちらか他のプレイヤーとの手持ち金の差が開かない行動を選択しており，プレイヤーエージェント A の ϵ が高い場合には，手持ち金が拮抗する状態になりやすいといえる．各プレイヤーの手持ち金が拮抗した状態は重要な意味を持つ．拮抗的な状態は行動選択の結果，順位などの変化が起きやすくユーザにとって楽しみやすいものであると考えられる．前述のプレイヤーエージェント A がユーザに置き換わった場合に，どのプレイヤーに対して攻撃的行動を選択するのかという選択によりユーザに対して過度に攻撃を行わないエージェントや，ユーザの行動選択を手助けするようなエージェント環境の実現の可能性があるといえる．

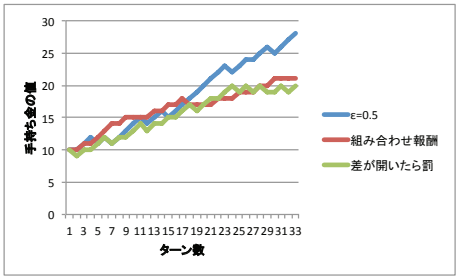

図 6: 順位に応じた報酬 (ϵ=0.5)，報酬の組み合わせ，差が開いた際の報酬にもとづいたプレイヤーエージェントの各プレイヤーの手持ち金の推移 (1 ゲーム)

6. まとめと課題

多人数戦略すごろくゲーム環境の設計と実装を行い，エージェントが学習に用いる複数の異なる評価関数を用意し強化学習を行った．順位に応じた評価関数を与えたエージェントの行動選択に用いる ϵ を工夫することで異なるレベルのユーザプレイヤーを擬似的に再現した．擬似的なユーザプレイヤーを参加させたゲーム環境において対戦相手として事前に異なる評価関数によって学習させたプレイヤーエージェントを設定し，どのようなプレイヤーが存在するゲーム環境がユーザにとって楽しみやすい環境であるかを考察した．

差が開いた際に負の報酬を与える評価関数を用いたエージェントを参加させたゲーム環境において，ユーザプレイヤーがゲームの状態に対してどのような評価を持つかを比較した．

参考文献

[Hu 15] Hu, Y., Gao, Y., and An, B.: Learning in Multi-agent Systems with Sparse Interactions by Knowledge Transfer and Game Abstraction, in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pp. 753–761, Richland, SC (2015), International Foundation for Autonomous Agents and Multiagent Systems

[Sutton 98] Sutton, R. S. and Barto, A. G.: *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition (1998)

[Wicke 15] Wicke, D., Freelan, D., and Luke, S.: Bounty Hunters and Multiagent Task Allocation, in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pp. 387–394, Richland, SC (2015), International Foundation for Autonomous Agents and Multiagent Systems

[杉浦 16] 杉浦 生隼, 福田 直樹：鉄道経営ゲームにおける思いやりある協調的動作の実現, 情報処理学会第 78 回全国大会講演論文集 (2016), 6M-05

[鈴木 01] 鈴木 麗璽, 有田 隆也：N 人版繰り返し囚人のジレンマゲームにおける空間的局所性の影響とその進化, 電子情報通信学会技術研究報告. AI, 人工知能と知識処理, Vol. 101, No. 66, pp. 39–45 (2001)