6M-05

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 *

杉浦生隼 † 福田直樹 ‡

静岡大学情報学部

1 はじめに

近年コンピュータゲームにおけるAIと人間との関わりは，一層増している[1]．エージェントの学習メカニズムを用いて複数のソフトウェアエージェントの協調的動作の実現も進みつつある[3]．本研究では，鉄道経営ゲームを模した環境においてゲームの特徴を学習したエージェントが，行動の価値を獲得し過度に他のプレイヤーを攻撃しないような協調的動作の実現を目指す．エージェントの学習に3種類の異なる報酬を与え，ゲーム環境上でエージェントがどのような行動選択の価値を学習可能であるかについて考察する．

2 本研究における鉄道経営ゲーム

本研究における鉄道経営ゲームは，マルチプレイヤーのボードゲームである．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．

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．

3 行動価値の獲得

ここでは実験的なゲーム環境上における Q-Learning[2] 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．

ゲーム環境には5つの駅から生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

*Toward Omoiyari-driven Cooperative Behavior on Playing Railway Management Games

†Kihaya Sugiura, Faculty of Information, Shizuoka University, 432-8011, Hamamatsu, Japan

‡Naoki Fukuta, Faculty of Information, Shizuoka University, 432-8011, Hamamatsu, Japan

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．

順位変動 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．

順位に応じた報酬 学習エージェントの順位に応じた報酬を，1位を最大の報酬値とし，与える．

攻撃的行動の有無 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．

4 結果と考察

表2に，7000ゲーム間における学習エージェントの各駅の選択回数の割合を示す．黄駅に止まることも攻撃的な行動に至る一連の行動を構成していると考えられるため，攻撃的な行動の選択の割合と合わせて，黄駅に止まる行動についても着目する．

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獲得され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にもかかわらず，攻撃的行動を選択する割街が3パーセントと，およそ4分の1程度となっている．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ほぼ同じである．

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

評価関数	青駅	赤駅	黄駅	攻撃的行動
順位変動	50.0%	24.2%	15.1%	12.1%
現在の順位	10.6%	66.6%	21.2%	3.0%
攻撃的行動の有無	21.2%	42.4%	34.8%	3.0%

5 まとめ

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ュ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3種類の評価関数を用意し，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．

参考文献

[1] 情報処理学会:コンピュータ将棋プロジェクトの終了宣言，http://www.ipsj.or.jp/50anv/shogi/20151011.html

[2] Sutton, R.S. and Barto, A.G.: Reinforcement Learning:An Introduction, MIT Press, Cambridge, 1998

[3] Drew Wicke, David Freelan, and Sean Luke. 2015. Bounty Hunters and Multiagent Task Allocation. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15).387-394.