1F2-2

人狼における強化学習を用いたエージェントの設計

Design of Agent in“ Are you a Werewolf? ” using Reinforcement Learning

梶原 健吾[*1] Kengo Kajiwara
鳥海 不二夫[*1] Fujio Toriumi
稲葉 通将[*2] Michimasa Inaba

[*1] 東京大学 The University of Tokyo
[*2] 広島市立大学 Hiroshima City University

In recent years, advances in artificial intelligence techniques in the complete information game are remarkable. Followed by Othello and chess, even in Shogi artificial intelligence has come to win the man of professional. On the other hand, there are few researches on artificial intelligence of incomplete information game, and there is room for development. Therefore it is considered to be one of the new targets of artificial intelligence. In this paper, we design the agent in “Are you a Werewolf?”, which is one of the incomplete information game, using reinforcement leaning. There are various strategies in “Are you a Werewolf?”, and strategies are different by those who play, so we design the agent that can take the appropriate action against a variety of opponents.

1. はじめに

近年，完全情報ゲームにおける人工知能の技術の進歩は目覚ましく，オセロやチェスに続き，将棋においても人工知能が人間のプロを相手に勝利するようになった．一方で，不完全情報ゲームの人工知能についての研究は少なく発展の余地があり，人工知能が目指すべき新たな目標の一つであると考えられる．

鳥海らは，[鳥海 14]にて不完全情報ゲームである“人狼”[稲葉 12],[大澤 00]を人工知能がプレイできるプラットフォームを構築した．また，梶原らは[梶原 14]にて別の“人狼”プラットフォームにおいて，ゲーム内の会話内容やプレイヤー数等を元に，発言内容，能力者の行動等を強化学習し，最適戦略の解析を行った．その結果，不完全情報ゲームにおいても有意な戦略を学習によって得ることが可能だと示した．しかし，[梶原 14]で利用された手法では，学習の際に用いた対戦相手が単一であり，学習結果はその対戦相手のみに最適化されたものである．そのため，学習時に用いていない新たなプレイヤーと対戦する際には最適でない行動を取る可能性がある．実際に“人狼”を遊ぶ際には，人によって用いる戦略が様々であるため，エージェントは多様な相手に対応できる必要性がある．

そこで，本研究では，[鳥海 14]で新たに構築された“人狼”のプラットフォームにおいて，[梶原 14]の手法で学習を行い，新しいプラットフォームにおける学習可能性を確かめる．また，[梶原 14]の学習手法を拡張することで，新たなプレイヤーに対しても最適な行動を取ることができる学習エージェントの設計を行う．

2. 学習手法

2.1 Q 学習

本研究では，エージェントの学習方法として Q 学習を用いた．Q 学習[三上 00]とは，ある状態 s_t において選択可能な行動を a_t としたときに，各行動の有効性を表す行動価値関数 $Q(s_t, a_t)$ (以下，Q 値)を評価する手法である．

Q 学習では繰り返し行うエピソードの各ステップにおいて，状態 s_t における行動 a_t を方策 $\pi(a_t, s_t)$ に従って選択し，その Q 値を更新していく．本研究におけるエピソードとは，1 回の人狼ゲームのことであり，状態 s_t は会話や投票結果等により表される t 日目の状況を指し，行動 a_t は発話内容の決定や占い対象の選択等を指す．行動 a_t を取ったときに，その行動によって得られる報酬 r_{t+1} とその行動によって移行した次の状態 s_{t+1} を観測し，それらをもとに $Q(s_t, a_t)$ を以下のように更新する．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

この式における α は学習率，γ は割引率と呼ばれ，それぞれ 0 以上 1 以下の値をとる．状態 s_t が終端状態，すなわち“人狼”においては人間側か人狼側のどちらかが勝利した状態に移行すると，そのエピソードを終了し，さらにエピソードを繰り返す．

2.2 人狼における Q 学習の適用

(1) 状態

人狼は不完全情報ゲームであり各プレイヤーの役職は公開されないため，会話内容や投票結果からは，複数の役職の組み合わせが考えられる．すなわち，各プレイヤーは現在の状態 s_t を正確に知ることはできない．

そこで，本研究では複数の状態を同時に扱うことにした．例えば，「agentA と agentB が占い師だとカミングアウト(以下，CO)し，agentA は agentC を人間だと判定し，agentB は agentC を人狼だと判定した．」という状況においては，表 1 に挙げた 9 パターンの役職の組み合わせが考えられるので，それぞれのパターンを一つの状態 s_t とみなし，現在の状態はこれらのいずれかであるとする．

(2) 行動

人狼における Q 学習の行動として以下の 3 項目を扱った．

- 能力者の CO 条件の選択

連絡先：梶原健吾，東京大学工学系研究科システム創成学専攻，〒113-8656 東京都文京区本郷 7-3-1 工学部 8 号館 526，TEL：03-5841-6991，E-mail：kajiwara@crimson.q.t.u-tokyo.ac.jp

表 1. あり得る役職パターン

パターン	agentA	agentB	agentC
1	占い師	狂人	人間
2	占い師	人狼	人間
3	狂人	占い師	人狼
4	狂人	人狼	人間
5	狂人	人狼	人狼
6	人狼	占い師	人狼
7	人狼	狂人	人間
8	人狼	狂人	人狼
9	人狼	人狼	人間

占い師, 霊能者, 人狼, 狂人はゲーム開始時に自分の役職を CO する条件を以下に示す要素の組み合わせから選択する.

- ゲームがn日経過(0<n<6)
- 能力で人狼を発見
- 人狼だと占われた
- 対抗能力者が出てきた
- 自分に投票するプレイヤーが多数

- 人狼, 狂人の騙り役職の選択

人狼と狂人はゲーム開始時に, 自分が騙る役職を村人, 占い師, 霊能者から選択する. また, 人狼は他の人狼が各役職を騙った際に自分が騙る役職も村人, 占い師, 霊能者から選択する.

- 投票, 占い, 護衛, 襲撃の対象選択

投票や占い等の対象を表 2 に当てはまるプレイヤーから選択する.

学習後の行動の選択方法は,

$$Q'(a_t) = \sum_{s_t} Q(s_t, a) \times (s_t \textit{が現れる頻度})$$

とした時の $Q'(a_t)$ が最大となる a_t で与えるものとする.

(3) 1 ゲームの学習方法

本研究では, 以下に示す流れで Q 学習を行う.

- ゲーム終了時に, 前プレイヤーの役職を公開し, 各日の真の状態 s_t を定める.
- 真の状態 s_t 各々において, ゲーム中に選択した行動 a_t の $Q(s_t, a_t)$ を更新する. 報酬 r_t は最終日の状態 s_t のみに与え, 自軍が勝った際は 100, 負けた際は 0 を与える.

表 2. 行動の対象となるプレイヤーの種類

占い師
霊能者
占い師を騙る人狼
占い師を騙る狂人
霊能者を騙る人狼
霊能者を騙る狂人
人狼だと判定を出されたプレイヤー
人間確定のプレイヤー
襲撃されたプレイヤー(人狼の偽占いの対象)
処刑されたプレイヤー(人狼の偽占いの対象)
上記以外のプレイヤー
ランダム

2.3 Q 学習の反復的利用

[梶原 14]では, 全プレイヤーが学習データ(Q(s_t, a_t)の集合)を共有していたが, 本研究における新たな手法では全プレイヤー別々の学習データを持ち, 以下の流れで学習を行う.

- 5 つの学習を行う環境を用意し, 11 人プレイで 100 万回プレイを行い, 学習を行う. (11×5 個の学習データが生成)
- 各環境において, 人間側と人狼側の勝率を算出する.
- 人間側の勝率が最も低い環境における, 人間側の学習データを初期値に戻す. 人狼側についても同様.
- 5×11 人のプレイヤーをランダムに入れ替えて, 新たに 5 つの環境を作成し, 1に戻る.

この行程を 3 回繰り返した. 各環境におけるプレイヤーの行動に変化をつけるため, 最初の行程では Q 学習における学習率 α を 0.5 と極めて高い値とした. 2 回目は α =0.3, 3 回目は α =0.1 とした.

3. 結果

今回用いる“人狼”プラットフォームにおける学習可能性を検証し, また, [梶原 14]の手法によるプレイヤーO と, 新たな手法によるプレイヤーN との強さを評価するため以下の実験を行った.

- 無学習状態のプレイヤー(以下, ランダムプレイヤー) 10 人の中に, プレイヤーO またはプレイヤーN を1体入れてプレイさせ, そのプレイヤーが各役職になった場合における人間側の勝率を評価する.

この実験結果を表 3 に示す. 勝率は全て人間側の勝率を表している. また, 役職が無しの項目は, プレイヤーO もプレイヤーN も入れずに, 11 人プレイを行った場合の村人の勝率を表している.

表 3 において, ランダムプレイヤーに対して全ての役職についてその役職側の勝率が上昇していることから, 学習はできていると判断出来る. しかし, 表 3 において, プレイヤーO, プレイヤーN の各役職時の勝率にはほとんど差が現れなかった.

新しい手法では, 学習時に様々な学習結果を持ったプレイヤーと対戦させたが, 対戦相手個々の行動を Q 学習における状態として認識することが無かった. そのため, 新たなプレイヤーと対戦したときに, 学習結果がうまく利用されなかった可能性が考えられる.

4. 結論

本研究では, 新たな“人狼”プラットフォームにおいてプレイヤーの強化学習を行った. また, [梶原 14]の学習手法を拡張し, 新たなプレイヤーに対して最適な行動を取る学習プレイヤーの設計を行った. 結果として, 新たな“人狼”プラットフォームにおいても強化学習が可能であることを示した. しかしながら, [梶原 14]の学習手法に対して優位なプレイヤーを作成することはできなかった.

今後の課題としては, 学習環境において必要な情報を Q 学習における状態に取り入れることが挙げられる. 必要な情報の選定はオートエンコーダ等を利用することで可能になると考えられる.

表 3. ランダムプレイヤーとの対戦時の勝率

	プレイヤーO	プレイヤーN
学習無し	36.2%	
狩人	49.6%	49.4%
占い師	50.5%	48.9%
霊能者	52.6%	50.5%
狂人	30.7%	33.3%
人狼	26.6%	27.4%
村人	48.4%	48.8%

参考文献

[鳥海 14] 鳥海不二夫, 梶原健吾, 大澤博隆, 稲葉通将, 片上大輔, 篠田孝祐:人狼知能プラットフォームの開発, 日本デジタルゲーム学会, 2015.

[梶原 14] 梶原健吾, 鳥海不二夫, 大澤博隆, 片上大輔, 稲葉通将, 篠田孝祐, 西野順二, 大橋弘忠:強化学習を用いた人狼における最適戦略の抽出, 情報処理学会, 出版社 ,2014.

[稲葉 12] 稲葉通将, 鳥海不二夫, and 高橋健一:“人狼ゲームデータの統計的分析,” in ゲームプログラミングワークショップ 2012 論文集, 2012, 144-147.

[大澤 00] 大澤博隆:“コミュニケーションゲーム「人狼」におけるエージェント同士の会話プロトコルのモデル化,” in HAI シンポジウム 2013, 2013, 122-130.

[三上 00] 三上貞芳・皆川雅章:強化学習, 森北出版株式会社, 2000.