情報処理学会論文誌 Vol.60 No.10 1728–1737 (Oct. 2019)

Q学習と役職推定に基づく人狼知能エージェントの作成

萩原 誠[1,a)] 伊藤 孝行[1,b)] アーメッド ムスタファ[1,c)]

受付日 2019年1月25日, 採録日 2019年7月3日

概要：本研究は，不完全情報ゲームである人狼ゲームにおいて，参加プレイヤ数が多く，ルールベースに基づくエージェントが対戦相手の場合を考慮した，強化学習を用いた人狼知能エージェントを作成する．エージェントによる人狼ゲームの大会が，人狼知能大会として 2015 年から実施されている．人狼知能大会では，エージェントの役職が人狼である場合の全体の勝率が低く，エージェントが人狼である場合の性能の向上が必要である．本研究では，エージェントの役職を人狼に固定して，役職の推定と Q 学習を併用したエージェントを提案する．本手法の有効性を参加プレイヤが，11 人の人狼（11 人村）および 15 人の人狼（15 人村）において過去の人狼知能大会に参加したエージェントと対戦させることで評価する．11 人村は人狼が 2 人，15 人村では 3 人であり，役職に占い師，霊能者と騎士が存在するためゲームが複雑になる．11 人村および 15 人村において既存のエージェントを相手とした場合に本手法が有効であることを勝率により示すことで，本手法の有効性が確認された．

キーワード：マルチエージェントシステム，人狼知能，強化学習

Werewolf Game Agent Using Q-learning and Estimation of Role

Makoto Hagiwara[1,a)] Takayuki Ito[1,b)] Ahmed Moustafa[1,c)]

Received: January 25, 2019, Accepted: July 3, 2019

Abstract: This paper focuses on construction of agent using reinforcement learning (RL) in large scale Werewolf Game with existing rule-based agents. AIwolf contest was held from 2015. AIwolf means the agent which play Werewolf Game. The win rate in this contest showed that it is necessary for AIwolf to improve performance of playing role of werewolf. In this paper, we propose a model that using RL for construction of strategy and estimating role in Werewolf Game. We showed through evaluation experiment that our approach has effectiveness of performance in large scale Werewolf Game with existing rule-based agents.

Keywords: multi-agent system, AIwolf, reinforcement learning

1. はじめに

近年，囲碁 [1] や将棋 [2] といった完全情報ゲームではコンピュータプログラムが人間の勝率を上回ることが示されてきた．完全情報ゲームとはすべてのプレイヤが行動するにあたり自分が現在どの状態にいるかを知覚しているゲームのことである．一方で，不完全情報ゲーム [3] を対象にした研究も行われている．不完全情報ゲームとは，「プレイヤごとに得られるゲームの状態に関する情報が部分的で不完全なゲーム」[3] である．人狼ゲームは不完全情報ゲームの 1 つである．人狼ゲームでは各プレイヤは人狼陣営と村人陣営の 2 つのチームのいずれかに所属しており，プレイヤを多数決によりゲームから取り除き，残るプレイヤが人狼陣営，または村人陣営のどちらかの陣営の人数が一定の条件を満たすと勝敗が決まる．人狼ゲームではゲーム中に発話が行われ，発話がゲームにおいて最も重要な要素である多数決に大きな影響を及ぼす．エージェント [4] が人狼ゲームを行うには発話と行動に関する戦略を学習する必要がある．エージェントとは「ある環境の中で自律的に，与えられた目的を達成するために行動するコンピュータシステム」[4] である．

[1] 名古屋工業大学
Nagoya Institute of Technoloy, Nagoya, Aichi 466–8555, Japan
[a)] hagiwara.makoto@itolab.nitech.ac.jp
[b)] ito.takayuki@nitech.ac.jp
[c)] ahmed@nitech.ac.jp

人狼ゲームでは各プレイヤの行動がゲームに大きく影響を与えるため,「他者による行動の影響」と「自分の行動が他者に与える影響」の 2 点を考慮した発話と行動が要求される．他者の行動の影響として，占い師の行動が，全員の投票行動に影響する例を示す．占い師は，プレイヤの 1 人が人狼であるかどうかを知る能力を持っている．したがって，占い師は，夜にあるプレイヤが人狼であることを知り，次の日の議論でそのプレイヤが人狼であることを他のプレイヤに伝えることができる．占い師に人狼だといわれたプレイヤは，他のプレイヤからの投票の対象になりやすい．すなわち，占い師の行動が，そのプレイヤに投票を集中させる可能性が高まる．一方，人狼であるといわれたプレイヤへの投票は，そのプレイヤ自身が村人側に協力的なことを伝え，自分への投票数を変化させようとする．これは村人側に協力的であるとされるプレイヤは投票されにくいためである．自分への投票数の変化は，自分の行動が，他者に与える影響の結果である．以上のように人狼ゲームでは他者の行動の影響，および行動による他者への影響を考慮する戦略が要求される．「他者による行動の影響」，および「自分の行動が他者に与える影響」の 2 つの影響を考慮する戦略をエージェントに学習させることは，相手の行動変化を考慮するうえで有効である．

2019 年現在，国内では，人狼知能大会 [5] と呼ばれる人狼知能エージェントの競技会が行われている．本研究では，人狼ゲームを行うエージェントのことを人狼知能エージェントと呼ぶ．人狼知能大会では 15 体のエージェントまたは 5 体のエージェントで人狼ゲームを行いその勝率を競う．国内から様々な戦略を持つ人狼知能エージェントが人狼知能大会に集まり，対戦が行われている．人狼知能大会は人狼知能プロジェクト [6] の一環として行われており，対戦には人狼知能プラットフォーム [7] が使用されている．

2018 年の大会である第 4 回人狼知能大会において人狼側の勝率は 3 割弱である．オンライン上で人狼ゲームを行うプラットフォームである人狼 BBS [8] の結果をもとに人狼ゲームの勝率と各役職の寄与を調査した稲葉ら [9] の研究によると人狼の勝率は回帰的に 5 割となると示唆されるため，人狼側の勝率が 3 割弱であることは低い値である．したがって，役職が人狼である場合の人狼知能エージェントの性能は，役職が人狼以外である場合の人狼知能エージェントの性能に比べて劣る．そこで人狼であるときにおける人狼知能エージェントの性能の向上が必要である．

プレイヤの役職を推定する，役職推定に関する既存研究では，深層学習を用いた大川ら [10] の研究，SVM [11] を用いた梶原ら [12] の研究がある．

人狼ゲームにおいて人狼が判明する可能性が上がることは勝敗に直結する．なぜなら人狼ゲームの勝利条件は，人狼の生存人数で定義されているからである．役職が人狼の場合はなるべく推定される行動をとらない戦略をとる必要がある．堂黒ら [13] の研究では，役職を推定するための特徴量に過去の投票履歴を利用することで，推定の精度を向上させている．このように投票行動を例として，プレイヤの行動は役職を推定するための情報を与えるため，人狼は自分の陣営の利益となる行動だけをとるべきではない．

以上より，推定による人狼の判明を避けるために，村人を模倣する行動と人狼の利益になる行動のバランスを調整した戦略が必要である．村人を模倣する行動の例としては，占い師の能力を使用した結果により人狼であると判定を受けたプレイヤに投票することで，村人に協力的なプレイヤを装い村人を模倣すること，人狼の利益になる行動の例としては，人狼以外のプレイヤに投票することで，人狼が生存する可能性を向上させることがあげられる．村人を模倣する行動と人狼の利益になる行動のバランスを調整した戦略をルールベースで実装することは以下の問題点がある．

- 人狼ゲームの参加人数の増加にともない，状態空間が指数関数的に増加するためルールベースによる実装のコストが高まる．
- ある行動が正しい判断であるかどうかは，判別を行うゲームの盤面を詳細に評価する必要があるため，ルールベースで行うのが困難である．

したがって，ルールベースによる記述でなく，エージェントに自ら戦略を学習させる必要がある．戦略の系列を学習する手法として強化学習が適している．加えて，役職が人狼である場合におけるエージェントの行動学習は，少数派の集団が議論などにより，多数派の集団を誘導する必要がある．そのため，少数派である人狼陣営の行動学習は，少数派の集団を有利にすることを目的としたエージェントによる行動の学習を，会話を含んだゲームにおいて確立することにつながると考える．本研究では少数派に人狼陣営が該当し，多数派に村人陣営が該当する．また，本研究における会話はプロトコルと呼ばれる簡易言語により行われるため，自然言語での議論と比べると非常に限定された状況である．本研究では強化学習によりエージェントが自ら，村人の行動の模倣と人狼の利益になる行動のバランスが調整された戦略を学習することを目指す．また本研究では，既存の人狼知能エージェントを対戦相手とした 11 人，および 15 人の人狼ゲームにおいて学習結果が有効であることを勝率を用いて確認する．勝率を用いて，現状のルールベースを用いた手法を相手とし，参加人数および，役職が人狼であるプレイヤの数が増加してゲームが複雑になった場合においても強化学習が有効であることを示す．11 人はレギュレーションで指定された人数ではないが，役職が人狼であるプレイヤの数が 2 人，および 3 人のどちらの場合も検証するために 11 人の場合も実験対象とする．どちらの場合についても 1 体のエージェントだけが学習を行う．

論文構成は 2 章で関連研究についてとりあげる．3 章で人狼ゲームの説明，人狼知能プラットフォーム，および人

狼知能プロトコルについて述べる．4 章で本論文で提案する人狼知能エージェントの手法について述べる．5 章で過去の人狼知能大会に参加した人狼知能エージェントと評価実験を行い，勝率を比較する．6 章で実験結果に基づいて考察を行う．7 章では本研究のまとめと今後の課題を示す．

2. 人狼知能

2.1 人狼ゲーム

人狼ゲームは会話と推理を中心にしたパーティーゲームである．プレイヤはそれぞれが村人と村人に化けた人狼となり，他のプレイヤと交渉して相手の正体を探る．人狼は自分の正体を隠しながら行動する．つまり村人陣営と人狼陣営の 2 つのチームに分かれる．ゲームは半日単位で進行し，昼には全プレイヤの投票により決定された人狼容疑者 1 人の処刑が行われ，夜には，人狼による村人の襲撃と，能力を持つプレイヤによる能力の使用が行われる．**表 1** に役職ごとの能力を示す．襲撃または処刑により死亡したプレイヤは以後ゲームに参加できない．すべての人狼を処刑することができれば村人陣営の勝ち，生存している人狼と同数まで村人を減らすことができれば人狼陣営の勝ちとなる．基本的なゲームの要素はプレイヤが秘密に割り振られた村人，人狼，占い師などの役職を互いに推理することである．しかし推理に必要な手がかりは一部の役職で断片的に知らされるほかは，他のプレイヤの役職を，プレイヤ同士のやりとりから読み取らなければならない．よって，いかに相手を説得するか，あるいは巧妙に騙し続けるかの会話による駆け引きが重要な要素となる [8]．

2.2 人狼知能プロジェクト

人狼知能プロジェクト [6] とは「人間と自然なコミュニケーションをとりながら人狼をプレイできるエージェントの構築」を目標としたプロジェクトである．人狼知能プロジェクト内において，人狼ゲームを行うエージェントの大会である人狼知能大会が行われている．人狼知能大会は 2018 年時点では「プロトコル部門」と「自然言語部門」に分かれており，プロトコル部門は簡易言語であるプロトコルによりすべての行動が決定される．一方で自然言語部門はゲーム中の議論に自然言語を用いる点でのみプロトコル部門と異なる．また参加人数に関しては，プロトコル部門は 15 人村と 5 人村，自然言語部門は 5 人村である．本研究はプロトコル部門の環境を利用する．

表 1 役職別の特殊能力
Table 1 Role ability.

役職	役職の能力
占い師	毎日夜に，生存しているプレイヤから 1 人を選択し，人狼であるかを知ることができる（占い）．
霊能者	毎日夜に，昼に処刑されたプレイヤが人狼であるかを知ることができる（霊能）．
騎士	毎日夜に，生存しているプレイヤから 1 人を選択し，人狼の襲撃から護衛できる（護衛）．襲撃先と重なれば，人狼の襲撃は失敗し，襲撃先のプレイヤはゲームから取り除かれない．
人狼	毎日夜に，生存しているプレイヤから 1 人を選択し，襲撃できる（襲撃）．
村人	能力を持たない村人側のプレイヤ．
多重人格	能力を持たない人狼側のプレイヤ．

2.3 人狼知能プラットフォームとプロトコル

ゲーム中の会話

人狼プラットフォームにおける日中の対話はターン制を導入している．以後，夜に行われる人狼の対話である囁きと区別するために対話と呼ぶ．囁きは人狼のみが行い，ターン制である．各プレイヤがターンごとに 1 回発言するチャンスがあり，各プレイヤの発言はまとめて他のプレイヤに送られる．順番はランダムであるため同一ターンの発言順に意味はない．発言は 1 日に 10 回まで行うことができる．ただし発言の見送りを意味する「Skip」とその日の議論の終了を希望する「Over」と呼ばれるものは発言に含まれない．すべてのプレイヤが Over を選択するか，すべてのプレイヤが同時に 3 回連続で Skip を選択するとその日の昼のフェーズは終了する．昼のフェーズは最大 20 ターンである．

プラットフォームではプロトコルを用いて会話が行われる．対話，および囁きにおいて可能な，主な発話の内容は以下のとおりである．

- 真偽問わず，自分の役職を宣言する（以後，CO と呼ぶ）．
- 能力の使用結果を伝える．
- 過去の発話に同意，または非同意する．
- 他のプレイヤの役職を推測する．
- 投票する相手を宣言する．

投票

襲撃は生存している人狼のプレイヤの投票，処刑は生存しているプレイヤの投票によりそれぞれ多数決で対象を決定する．投票の結果，最多得票者が複数人いた場合は 1 度だけ再投票が行われる．再投票でも最多得票者が複数人いる場合は最多得票者からランダムに選択される．

3. 関連研究

3.1 強化学習と Q 学習

強化学習問題とは対象について不完全な知識しかなく，また対象への働きかけにより観測できることが変化する場合に最適な働き方の系列を発見する問題である．強化学習では働きかける主体をエージェントと呼び，働きかける対象を環境と呼ぶ．エージェントによる環境への働きかけを行動と呼び，エージェントの行動により環境に何らかの変

化が生じる．環境を構成する要素のうち，変化する要素を状態と呼ぶ．同じ行動でも環境の状態により生じる結果は変化する．環境における行動を統一的に評価する指標として，報酬と呼ばれる数値により行動の結果の良さを表現する．強化学習問題とは与えられた環境の中で行動を通して得られる報酬の総和を最大化する問題である．

本研究においては強化学習のアルゴリズムである Q 学習 [14] を利用する．Q 学習は強化学習アルゴリズムのうちの 1 つである．Q 学習とはある状態における行動の価値を示す値である Q 値を更新し，Q 値に基づいて行動をするアルゴリズムのことである．更新式を式 (1) に示す．Q 値は行動直後に与えられる即時報酬 R と遷移先の状態における Q 値の最大値を用いて更新する．α は学習率と呼ばれ，現在の Q 値をどの程度に重視するかを決定する．γ は割引率と呼ばれ，今後の行動で与えられる報酬である遅延報酬をどの程度評価するかを決定する．

$$Q(s_t,a_t) = (1-\alpha)Q(s_t,a_t) + \alpha(R(s_t,a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1})) \quad (1)$$

Q：ある状態に対する行動の価値
R：ある状態のある行動に対する報酬
s_t：現在の状態　a_t：現在の状態における行動
s_{t+1}：次の状態　a_{t+1}：次の状態における行動
α：学習率　γ：割引率

3.2 人狼知能における強化学習

人狼知能において強化学習を用いた既存研究は著者らが知る限り 4 件存在し，人狼知能プラットフォームを用いない 1 件と，人狼知能プラットフォームを用いた 3 件がある．人狼知能プラットフォームを用いない研究は，梶原ら [15] による Q 学習を用いた研究がある．人狼知能プラットフォームを用いない梶原らのエージェントは，すべての役職を学習対象としており，ゲーム内の発話や人数をもとに，「能力を使用する対象の選択方法」や「人狼陣営が騙る役職」を学習し，人狼ゲームにおいて強化学習が有効であることを示した．人狼知能プラットフォームを用いない梶原らの強化学習を用いたエージェントの研究は戦略の選択に $\epsilon - greedy$ [16] を用いている．$\epsilon - greedy$ は行動の選択時に確率 ϵ でランダムに行動の選択を行い，それ以外は最も利得の高い選択肢を選択するアルゴリズムである．本論文では，利得の最も高い行動を選択することは Q 値が最も高い選択肢を選ぶことと同じである．提案手法との大きな違いとして，プラットフォームに人狼知能プラットフォームを用いない点があげられる．

人狼知能プラットフォームを用いた，強化学習の 3 件の説明を行う．人狼知能プラットフォームを用いた梶原ら [17] の研究は人狼知能プラットフォームを用いない梶原らの Q 学習の手法を拡張することで人狼知能プラットフォームでの強化学習が可能であることを確かめた．人狼知能プラットフォームを用いた梶原らの研究では状態として「会話内容に基づいて推測したすべてのプレイヤの可能な役職の組合せ」，行動として「投票，占い，護衛，襲撃の対象選択」，「多重人格，人狼が騙る役職」，「CO をするタイミング」の 3 つを定義している．また行動の対象としては以下の種類が存在する．

- 占い師
- 霊能者
- 占い師を騙る人狼（または多重人格）
- 霊能者を騙る人狼（または多重人格）
- 人狼だと判定を出されたプレイヤ
- 人間確定のプレイヤ
- 襲撃されたプレイヤ
- 処刑されたプレイヤ
- 上記以外のプレイヤ
- ランダム

学習後の行動選択方法は式 (2) としたときの $Q'(a_t)$ が最大となる a_t で与えられるとする．

$$Q'(\alpha_t) = \sum_{s_t} Q(s_t,\alpha) \times (s_t\text{が現れる頻度}) \quad (2)$$

Q：ある状態に対する行動の価値
s_t：現在の状態　a_t：現在の状態における行動
α：学習率

稲葉ら [9] は騎士と占い師が村人側の勝利に大きく寄与していることを確認している．したがって騎士も行動対象に含めるべきであると考えられる．しかし提案手法との違いとして，人狼知能プラットフォームを用いた梶原らの研究は，行動の対象にゲーム中に存在する役職である騎士を考慮していないため，本研究は騎士の役職推定を行うことで騎士も行動の対象に選択することを可能にする．人狼知能プラットフォームを用いた梶原らの研究は参加人数が 11 人の人狼ゲームにおいて強化学習の有効性を示した．ただし提案手法との違いとして，対戦相手としてはランダムな行動を中心とするエージェントを相手にしているためルールベースで実装されたエージェントを相手にした場合の強化学習の有効性を示していない．Q 学習を利用した王ら [18] の研究は「投票，占い，護衛，襲撃の対象選択」の学習を目指した．状態としては会話から抽出した「CO した役職（宣言）」，「他のプレイヤを信用するか疑うか（態度）」，「確定した役職（事実）」を状態としている．対象の選択としては以下の基準を用いている．

- 一番疑わしいプレイヤを選ぶ．
- 一番疑われるプレイヤを選ぶ．
- 反対態度を持つ 2 人のプレイヤの中から 1 人を選ぶ．
- 発言が一番多いプレイヤを選ぶ．
- 発言が一番少ないプレイヤを選ぶ．

- ランダムにプレイヤを選ぶ．

人狼知能プラットフォームを用いた梶原らの研究，Q 学習を利用した王らの研究のいずれも行動に対してヒューリスティックを用いている点は共通している．

DQN [20] を用いた王ら [19] の研究は人狼知能の研究において初めて DQN を使用することで行動にヒューリスティックスを用いずに学習を行った．DQN を用いた王らの研究では，実験環境として以下の 2 種類の環境で学習を行った．

- プラットフォームに付随したランダムな行動を中心とするサンプルエージェント（A）
- 人狼知能大会の参加エージェント（B）

王らの DQN を用いた提案エージェントは，環境（A），および環境（B）のいずれにおいても一部の人狼知能大会の参加エージェントより高い勝率を示した．提案手法とは異なり，参加人数が 5 人という比較的簡易な場合に限定して，既存の手法に対しての有効性を示した．人狼ゲームは参加人数が増加すると，ゲーム日数や役職の組合せが増加する．これにより，ゲームがより複雑になるため，改めて強化学習の有効性を確かめる必要がある．また，王らの提案手法では行動の対象をインデックスによって決定しているが，インデックスによる行動選択を行うと，組合せがエージェント数に比例して指数関数的に増加する．エージェントの組合せ，および役職の組合せの両方を順列ありの組合せで考慮する必要があるため，学習時間の問題から人数の拡張に対応するのは困難であると考えられる．

4. 役職推定を併用した強化学習エージェント

4.1 概要

提案手法は強化学習を用いてエージェントの行動を学習させるとともに，役職推定を利用する．提案手法における役職推定の利用とは，エージェントの行動において何らかの役職を持つエージェントを対象にする場合に役職推定の結果で最もその役職である確率が高いエージェントを対象に行動させる形で役職推定のアルゴリズムを利用することである．**図 1** に手法の行動選択の概要を示す．図 1 の例では，(1) 現在の状態の Q テーブルのリストを参照し，

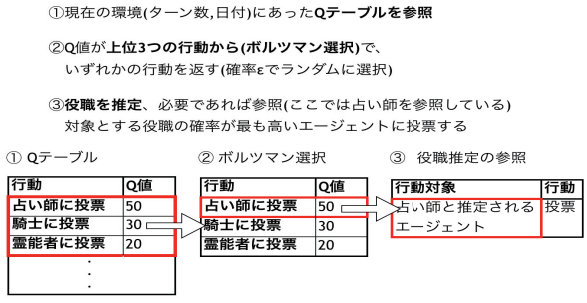

図 1 提案手法の行動選択の概要

Fig. 1 Overview of action decision in proposed model.

(2) 参照したリストから Q 値の最も高い 3 つの行動（占い師に投票，騎士に投票，霊能者に投票）を選び，その 3 つの行動からボルツマン選択で 1 つの行動（占い師に投票）を選択する．(3) 占い師を参照する必要があるため，役職推定を用いて占い師を推定し，占い師と推定されたエージェントに投票する．本研究では役職推定のアルゴリズムとして過去の人狼知能大会で使用されたナイーブベイズを利用する．本研究で用いるナイーブベイズのアルゴリズムは人狼知能エージェント Cndl 内で利用されているナイーブベイズのアルゴリズムを参考にしている．Cndl 内で利用されているナイーブベイズのアルゴリズムは日付，ターンなどの各状態において，採択した行動の頻度をもとにして各エージェントの役職の確率分布を更新する．ゲーム終了時に各エージェントのゲームで採択した行動により確率尤度を更新する．確率尤度は過去のすべてのゲームで採択した役職ごとにおける行動の頻度で表現される．

4.2 報酬の設計

提案手法ではゲームの途中でも報酬を受け取る．

【生存人数の変動について】

襲撃または処刑により生存人数が変動した場合は以下の報酬を与える．

- 人狼の人数が減少した：−2
- 村人の人数が減少した：+1

死亡したプレイヤの役職，死亡した日付の 2 点に応じて追加で報酬を与える．

- 人狼が取り除かれた：$-5 \times (7 - \text{日付})$
- 占い師が取り除かれた：$5 \times (7 - \text{日付})$
- 霊能者，または騎士が取り除かれた：$4 \times (7 - \text{日付})$
- 多重人格が取り除かれた：$-3 \times (7 - \text{日付})$

村人側の役職は占い師 ＞ 霊能者 ＝ 騎士（以下，この 3 つを能力持ち）の順で重要であると考え，人狼側の役職である多重人格の死亡には負の報酬を与える．能力持ちを序盤に取り除くことは，人狼に利益があるため日付による重み付けを行う．なぜなら，能力持ちの生存日数が長いほど，人狼の正体が判明するリスク，および護衛による襲撃失敗のリスクが上昇するからである．

一方で人狼は能力持ちに襲撃したいが，騎士による護衛成功が発生する状況も避けたい．したがって，襲撃が失敗した場合には負の報酬を与えることで役職持ちへの過剰な襲撃を抑制する．

- 襲撃に失敗した：−15

【ゲーム終了時の報酬】

ゲーム終了時に報酬は，終了時の村人，および人狼の人数と敗北した過去 1,000 ゲームまでの村人の人数により変化させる．人狼が勝利した場合には以下の報酬を与える．

- 自分が生存した：+80
- 生存した自分以外の人狼 1 体につき：+40

村人が勝利した場合には以下の報酬を与える．

- 20×(生存人数–敗北した過去 1,000 ゲームの村人の生存人数の平均)

人狼勝利時の評価については勝利に寄与する要因の評価は困難であるため，自分が生存して勝つ方が自分が死亡して勝利する場合より大きな報酬を得る．一方で村人の勝利時に関しては学習が進むにつれて村人の残り人数が変動するため，平均人数の変動に応じて評価を変動させる．生存人数の変動を考慮することにより，学習が進むにつれて村人の平均生存人数がより減少する状態においても報酬を適切に与えることを目指す．

4.3 行動選択時の確率

学習過程において探索と利用のトレードオフは重要である．ここでいう利用は探索により得た知識を利用することを意味する．提案手法では探索を促進するために $\epsilon-greedy$ 方策とボルツマン選択を併用した．まず確率 ϵ_{Act} でランダムで行動を選択する．それ以外の場合は Q 値の大きな順に 3 つの行動を候補として，ボルツマン選択を行う．ただしボルツマン選択における数値として最低値を e としている．つまり，候補として 1 より小さい Q 値を持つ行動を選んだ場合にボルツマン選択における数値は e として扱われる．また，このボルツマン選択とランダム選択の分岐を行う確率 ϵ_{Act} については該当する状態空間がそれまでに呼び出された回数により変動させる．確率 ϵ_{Act} の変動について以下に示す．

1 日目 1 ターン目の発話 $\max(0.1, \exp(-\frac{n}{10{,}000.0}))$
2 日目以降 1 ターン目の発話 $\max(0.1, \exp(-\frac{n}{2{,}000.0}))$
投票または襲撃 $\max(0.1, \exp(-\frac{n}{2{,}000.0}))$
2 ターン目以降の発話 $\max(0.1, \exp(-\frac{n}{500.0}))$

4.4 学習する行動の説明

学習する行動の要素は大きく 3 つに分かれる．

投票 毎日における投票先
襲撃 毎晩の襲撃先
発話 昼の議論において各ターンで行う発話の種類

各行動の対象としては以下の分類が存在する．

- 役職推定による，各役職である確率が最も高い者
- 占い判定を誰にも出されていない者
- 占い判定で人狼の判定を受けたことのある者
- 占い判定で人間の判定のみを受けている者
- 今まで人狼への投票数が一番多い者
- 最も多くのプレイヤから投票宣言を受けている者
- ランダム

行動の候補は 3 種類選ばれるが，どの行動も実行できない場合がある．たとえば，投票を行う状況で 3 パターンで選ばれた全員が死亡している場合は実行できない．行動の候補がすべて実行できないときは，発話の場合は何もせず，投票と襲撃の場合はランダムに対象を選択する．

4.5 事前学習について

既存の人狼知能エージェントを事前学習なしに相手にすると，初期段階における学習に失敗する，または過剰に時間が必要になる可能性がある．したがって，提案手法ではサンプルエージェントを相手に事前に学習を行う．サンプルエージェントはルールベースであり，相手の発話による投票先への影響は能力使用の結果の発話のみにしか影響されない．したがって，本手法では能力使用に関する発話は行わないため，サンプルエージェントを相手にした場合は発話の選択に関する学習が進行しない．そこでサンプルエージェントで投票と襲撃に関して学習させてから，既存の人狼知能エージェント群を相手に発話の選択に関する学習と，投票と襲撃に関する追加の学習を行う．問題点としてはサンプルエージェントの場合には学習できない状態空間も存在する点が問題である．例としてはサンプルエージェントが相手の場合は霊能者と占い師の CO の人数が 1 人ずつとなるケースが存在しないため学習を行うことはできない．

5. 実験

5.1 使用するエージェント

【評価実験用の学習エージェント】

提案手法を導入したエージェントに学習データを与えるため，11 人村ではサンプルエージェントと 200 万回の対戦を行った後に，過去の人狼知能大会決勝（またはそれに準ずる能力を持つエージェント）に参加した 10 種類のエージェント（以下，大会決勝エージェント）と 5 万回の対戦を行う．15 人村ではサンプルエージェントと 1,000 万回の対戦を行った後に，大会決勝エージェントにサンプルエージェントを 4 体加えたうえで，15 万回の対戦を行う．使用するエージェントは下記のとおりである．

- サンプルエージェント
- Cndl（第 3 回人狼知能大会）
- rsatio（人狼知能プレ大会@GAT2018 優勝）
- Romanesco（第 4 回人狼知能大会）
- AITKN（第 3 回人狼知能大会）
- wasabi（第 3 回人狼知能大会）
- aplain（第 3 回人狼知能大会）
- kasuka（第 3 回人狼知能大会）
- M-cre（第 3 回人狼知能大会）
- f6wbl6（第 4 回人狼知能大会）
- Udon（第 3 回人狼知能大会）

【評価用エージェント】

評価実験として提案手法のエージェント（提案エージェント）の役職を人狼に固定したうえで大会決勝エージェントと対戦を行う．対戦は 1 万回を 10 セットの計 10 万回の

対戦を行う．エージェント Udon，エージェント Cndl を提案エージェントと同じく役職を固定したうえで大会決勝エージェントと対戦を行う．そして提案エージェント，エージェント Udon，およびエージェント Cndl の 3 体の勝率を比較する．実験は人狼知能プラットフォームを使用する．大会決勝エージェントは過去の人狼知能大会において決勝に進出したエージェント群のうち Java で記述された 10 種類のエージェントである．15 人村では，11 人村の条件に 4 体のサンプルエージェントを 4 体加えた 15 体とする．人狼知能大会に出場したエージェントは python で記述されたエージェント（以下，python エージェント）も存在するが，以下の理由から Java で記述されたエージェント（以下，Java エージェント）のみを対象に対戦を行う．

- python エージェントを利用する場合に必要な TCP/IP を用いた通信接続による対戦の場合は実行時間が非常に長くなる．

大会決勝エージェント（15 人村では，サンプルエージェントを 4 体追加する）に勝率を確認するエージェント（提案エージェント，Udon，または Cndl のいずれか）1 体を加えた 11 体のエージェントで対戦を行い，勝率を比較する．ただし，勝率を確認したいエージェント以外は，毎ゲームごとに生成するため学習を行うことはできない．したがって，学習を行うのは勝率を確認するエージェントのみである．

5.2 実験結果

本節では，提案エージェントの学習の過程を示した後に，評価実験で用いた各エージェント，および提案エージェントとの比較をした結果について述べる．

11 人村において，提案エージェントの学習前の勝率は 60%である．また，提案エージェントの学習後では勝率が 67.5%である．ただし，勝率 67.5%は探索のためのランダム行動を含んだ場合の勝率である．ランダム行動を除いた場合は勝率が 69.1%に上昇した．

サンプルエージェント 10 体を相手にした場合と大会決勝エージェントを相手にした提案エージェントの勝率を，事前学習ありの提案エージェント，および事前学習なしの提案エージェントのそれぞれの勝率を**表 2** に示す．

表 2 の提案エージェント（事前学習あり）は，11 人村において，サンプルエージェントを相手に事前学習を行い，大会決勝エージェントを相手に 5 万回対戦を行った．提案エージェント（事前学習なし）は大会決勝エージェントを相手に 5 万回対戦を行った．表 2 の勝率は事前学習ありの提案エージェント，事前学習なしの提案エージェントがそれぞれ大会決勝エージェントと 10 万回の対戦を実施した勝率，およびサンプルエージェントと 10 万回の対戦を実施した勝率である．

表 2 から，サンプルエージェントを相手に事前学習した方が事前学習しないよりも，大会決勝エージェント，およびサンプルエージェントとの対戦の勝率のどちらにおいても高い値が得られた．そのため，サンプルエージェント相手に学習した内容が大会決勝エージェント相手でも勝率への寄与が示唆される．

Cndl，Udon，および提案エージェントとの勝率の比較を**図 2** で示す．図 2 の縦軸は勝率である．左の青色（提案エージェント），橙色（Udon），および灰色（Cndl）のグラフはサンプルエージェントを対戦相手にした場合における勝率を示す．右の青色，橙色，および灰色のグラフは大会決勝エージェントを対戦相手にした場合における勝率を示す．提案エージェントは，いずれの対戦相手の場合も最も高い勝率を示した．Udon がサンプルエージェントを相手にした場合の勝率は，他の 2 種類のエージェントと比較すると低い結果である．理由として，Udon は村人側の行動を模倣する割合が多いため，他のエージェントに比べて人狼側の利益となる投票先，および襲撃先の選択が少ないことがあげられる．人狼が村人側の行動を模倣した場合も能力使用の結果以外にはサンプルエージェントの行動への影響がない．したがって，人狼側の Udon が村人側の行動を模倣すると勝率が低下する．

続いて**図 3** に，図 2 の結果から提案エージェント，および Udon の勝率について $p < 0.01$ で χ^2 検定を行った結果を示す．図 3 の縦軸は勝率の平均，横軸は Udon，および提案エージェントを示す．図 3 の χ^2 検定から Udon の勝率と提案エージェントの勝率は $p < 0.01$ であるため，有意差が認められた．勝率の平均については提案エージェントが 34.7%，Udon が 33.4%であり，ϕ 係数を効果量として計算した数値は 0.0139 である．しかしながら，本研究では 11 体のうち 1 体の行動が変化するだけであるため全体へ

表 2 提案エージェントの勝率（11 人村）
Table 2 Proposed model's win rate (11 players).

	サンプル	決勝エージェント
提案エージェント（事前学習あり）	69.1%	34.8%
提案エージェント（事前学習なし）	62.3%	31.5%

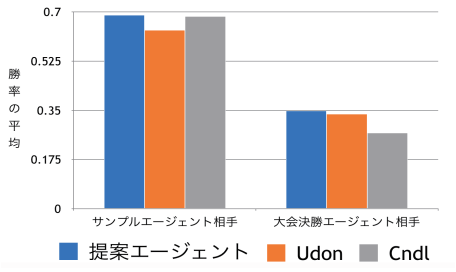

図 2 サンプルエージェント，および大会決勝エージェントとの対戦（11 人村）
Fig. 2 Result with sample agents and result with contest agents (11 players).

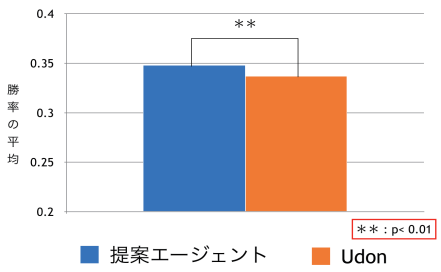

図 3 勝率と検定
Fig. 3 Win rate and test.

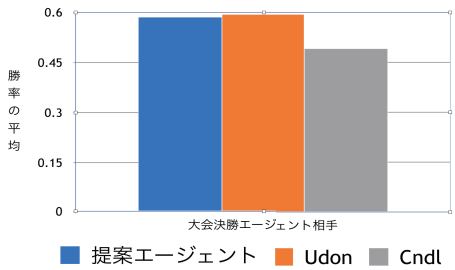

図 4 サンプルエージェント，および大会決勝エージェントとの対戦（15 人村）
Fig. 4 Result with sample agents and result with contest agents (15 players).

の影響が限られている．したがって勝率の変化においては数%の差に収まることが想定されるため，効果量が小さい値を示すのは妥当である．Cndl についてもゲームを行った回数が提案エージェントと同じであり，勝率が Udon よりも低いことから提案エージェントとの有意差が認められた．続いて，15 人村における勝率を示す．15 人村では勝率の変化は，サンプルエージェントを相手にした場合では 70.3%から 81.1%に変動した．ランダム行動を除いた場合に，サンプルエージェント相手に 10 万回の対戦を行った場合の勝率は 82.2%まで向上した．さらに 15 人村において，提案エージェント，Udon，Cndl の 3 つのエージェントの勝率を比較した結果を**図 4** に示す．縦軸と横軸の扱いは，図 2 の場合と同様であるが，対戦回数はそれぞれ 7 万回とする．15 人村の場合は，提案手法は既存エージェントよりも少し低い勝率を示した．既存エージェントは，レギュレーションの設定である 15 人に，ハードコーディングで対応しているため，その分性能が向上したと考えられる．

6. 考察

サンプルエージェントと対戦することは，襲撃，および投票により勝敗が決まることが多い．したがって役職持ちへの襲撃と護衛の成功のリスクを考慮したうえで襲撃先の選定を学習していることが推測される．またサンプルエージェント相手には投票先の学習をしたことも推測される．これらの投票先と襲撃先についての学習した内容が，既存のエージェントを相手にした場合でも有効であり，勝率に

表 3 人狼知能大会 2018 の勝率（全体）[6] 掲載データより作成
Table 3 Win rate (all) in AIwolf Contest 2018 (in Ref. [6]).

エージェント名	全体の勝率
Cndl	60.8%
Udon	60.2%
WolfKing	58.5%
Romanesco	58.4%
marky	58.3%
f6wbl6	57.9%
wasabi	57.9%
Litt1eGirl	56.7%
yskn67	56.5%
sonoda	56.1%
TRKOkami	55.3%
spicy2	55.3%
WordWolf	54.1%
cash	51.7%

反映されていると考えられる．しかしサンプルエージェントを相手にした場合では学習できない状態空間も存在する．たとえばサンプルエージェントを相手にした場合は CO の人数で一部遷移しない状態が存在する．占い師の CO 人数が 1 人の状態のような，サンプルエージェント相手に学習できないような状態に関しても事前学習を行うことができれば，既存のエージェントを相手にした場合の勝率がさらに上昇する可能性がある．また，発話の選択に関する学習では，大会決勝エージェントを相手に行うことで学習できると推測されるが今回の実験では検証することができなかった．同様にサンプルエージェントは発話の選択に関する学習を進めることはできなかった．発話選択における学習の有無を確認するため，実際に学習したエージェントを以下のように分けて勝率を比較した．

A 投票，襲撃，発話を学習した内容で選択
B 投票，襲撃を学習した内容で，発話はランダムで選択

A 群（69.1%）と B 群（69.0%）においてサンプルエージェントを相手にした場合における勝率に変動は見られなかった．

次に，人狼知能における勝率の 1%の差が重要であることを示す．**表 3** と**表 4** に示すのは人狼知能大会 2018 における勝率である．

全体の勝率は表 3 に示すとおり，全体のうち最下位のエージェントを除いたエージェントが 7%の間に勝率が収まる．15 人村の人狼は 11 人対 4 人のゲームであるため，勝率が同陣営のプレイヤの影響を受けることで各エージェントの勝率が近くなる．11 人の人狼も 8 人対 3 人のゲームであり，チームの人数の比率がほぼ同じであることから勝率も同じくらい影響を受けると考えられる．したがって，各エージェント間における性能の差は数%の勝率の中に現れる．そのため，1%という勝率の差は人狼ゲームにおいては大きな意味を持つ．つまり，人狼ゲームにおける勝率

表 4 人狼知能大会 2018 の勝率（役職が人狼）[6] 掲載データより作成

Table 4 Win rate (werewolf) in AIwolf Contest 2018 (in Ref. [6]).

エージェント名	人狼の勝率
Udon	34.9%
cndl	32.9%
wasabi	32.9%
Litt1eGirl	31.4%
yskn67	31.3%
spicy2	29.7%
sonoda	28.6%
f6wbl6	28.0%
Romanesco	27.8%
marky	27.7%
TRKOkami	27.5%
WolfKing	27.2%
WordWolf	23.2%
cash	22.0%

1%の差は人狼知能プラットフォームを用いた人狼知能においても重要である．以上より，評価実験の結果では提案エージェントが 11 人村で評価用エージェントに対し，勝率が 1%を超える差を示したことから，11 人村における提案エージェントの有効性が確認された．

一方で，15 人村の場合は，Udon の方が提案エージェントに対して少し高い成績を示した．比較手法である Udon は評価関数をベースとしてルールベースによるハードコーディングを 15 人村用に行っているため，本手法よりも高い性能を示したと考えられる．しかし Udon の評価関数は人数ごとにヒューリスティックスで設定する必要があり，人数の拡張ごとに評価関数を設定するのはコストが大きいと考えられる．一方で，本手法は人狼の人数が変化した場合に，個々の人狼を対象として選択する行動を追加する以外には，特に人数を特定したハードコーディングによる変更は必要がないという点で優れている．

7. 結論

本研究は，不完全情報ゲームである人狼ゲームにおいて，参加プレイヤ数が多く，ルールベースに基づくエージェントが対戦相手の場合の強化学習を用いた人狼知能エージェントを作成した．提案手法は人狼知能プラットフォームにおいて，強化学習により行動の選択における戦略の学習を行い，役職の推定として人狼知能大会において有力な手法であるナイーブベイズによる役職推定を使用した．本研究では評価実験として，11 人村，および 15 人村において過去の人狼知能大会に参加したエージェントを相手に，エージェントの役職を人狼に固定して対戦を行い，既存の手法に対して，11 人村では勝率に有意差があることを示し，15 人村においては，人狼の役職であるエージェントとして有力であるエージェントと僅差の勝率を示した．今後の課題として，人狼陣営の最大の利点であるゲーム開始時点に人狼であるプレイヤの情報を活かすために，役職が人狼であるプレイヤ全体として協調する行動，および撹乱する行動の学習方法を確立することがあげられる．また，既存エージェント群を相手にゲームを行う場合に 1 ゲームあたりの実行時間が長いため，学習におけるサンプル効率の向上を目指すことも課題である．

謝辞 論文執筆についてご助言をいただいた藤田医科大学の奥原俊助教に厚く感謝申し上げます．また，所属研究室の皆様，特に北川峻也さん，丹田尋さん，故 遠山竜也さん，芳野魁さんには研究の基本についてご教授いただき心より感謝いたします．実験で使用したプラットフォームおよびプロトコルを利用させていただいた人狼知能プロジェクトの皆様，および実験でエージェントを使用させていただいた人狼知能エージェント作成者の皆様に感謝の意を表します．

参考文献

[1] Silver, D., Schrittwieser, J., Simonyan, K., et al.: Mastering the game of Go without human knowledge, *Nature*, Vol.550, pp.354–359 (Oct. 2017).
[2] Silver, D., Hubert, T., Schrittwieser, J., et al.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, arXiv:1712.01815 (Dec. 2017).
[3] 西野哲朗：不完全情報ゲーム，情報処理，Vol.53, No.2, p.117 (2012).
[4] Wooldridge, M. and Jennings, N.: Intelligent agents: Theory and practice, *The Knowledge Engineering Review*, Vol.10, No.2, pp.115–152 (1995).
[5] 人狼知能大会，入手先 ⟨http://aiwolf.org/aiwolf_contest⟩（参照 2018-12-17）.
[6] 片上大輔，鳥海不二夫，大澤博隆ほか：人狼知能プロジェクト（〈特集〉エンターテイメントにおける AI）Project AI Wolf，*Journal of the Japanese Society for Artificial Intelligence*, Vol.30, No.1, pp.65–73 (2015).
[7] 鳥海不二夫，梶原健吾，大澤博隆ほか：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).
[8] 人狼 BBS まとめサイト，入手先 ⟨https://wolfbbs.jp/⟩（参照 2018-12-17）.
[9] 稲葉通将，鳥海不二夫，高橋健一：人狼ゲームデータの統計的分析，*The 17th Game Programming Workshop* (2012).
[10] 大川貴聖，吉仲 亮，篠原 歩：深層学習を用いて役職推定を行う人狼知能エージェントの開発，*The 22nd Game Programming Workshop* (2017).
[11] Cortes, C. and Vapnik, V.: Support-vector networks, *Machine Learning*, Vol.20, No.3, pp.273–297 (1995).
[12] 梶原健吾，鳥海不二夫，稲葉通将ほか：人狼知能大会における統計分析と SVM を用いた人狼推定を行うエージェントの設計，*The 30th Annual Conference of the Japanese Society for Artificial Intelligence* (2016).
[13] 堂黒浩明，松原 仁：ニューラルネットワークを用いた人狼推定における投票先情報の有効性評価，GAT2018 論文集，pp.1–4 (2018).
[14] Watkins, C. and Dayan, P.: Q-learning, *Machine Learning*, Vol.8, No.3-4, pp.279–292 (1992).

[15] 梶原健吾，鳥海不二夫，大澤博隆ほか：強化学習を用いた人狼における最適戦略の抽出，情報処理学会第 76 回全国大会 (2014).
[16] Auer, P., Cesa-Bianchi, N. and Fischer, P.: Finite-time Analysis of the Multiarmed Bandit Problem, *Machine Learning*, Vol.47, pp.235–256 (2002).
[17] 梶原健吾，鳥海不二夫，稲葉通将：人狼における強化学習を用いたエージェントの設計，*The 29th Annual Conference of the Japanese Society for Artificial Intelligence* (2015).
[18] 王　天鶴，金子知適：人狼ゲームエージェントにおける行動選択手法の比較，ゲームプログラミングワークショップ 2017 論文集，Vol.2017, pp.177–182 (2017).
[19] 王　天鶴，金子知適：人狼エージェントにおける深層 Q ネットワークの応用，ゲームプログラミングワークショップ 2018 論文集，Vol.2018, pp.16–22 (2018).
[20] Mnih, V., Kavukcuoglu, K., Silver, D., et al.: Human-level control through deep reinforcement learning, *Nature*, Vol.518, No.7540, p.529 (2015).

萩原 誠　（学生会員）

2019 年名古屋工業大学情報工学科卒業．2019 年より同大学大学院工学研究科情報工学専攻博士前期課程在学中．学士（工学）．強化学習，マルチエージェント強化学習，およびマルチエージェントシステムに興味がある．人工知能学会学生会員．

伊藤 孝行　（正会員）

2000 年名古屋工業大学大学院博士後期課程修了．博士（工学）．1999 年 JSPS 特別研究員．2000 年 USC/ISI 客員研究員．2001 年北陸先端科学技術大学院大学助教授．2003 年名古屋工業大学大学院助教授．2005 年ハーバード大学および MIT 客員研究員．2006 年より名古屋工業大学大学院産業戦略工学専攻准教授．2008 年 MIT 客員研究員．2009 年 JST さきがけ大挑戦型研究員．2014 年名古屋工業大学大学院産業戦略工学専攻/情報工学教育類教授，現在に至る．2011 年内閣府最先端・次世代研究開発プロジェクト代表研究者．2015 年名古屋工業大学大学院産業戦略工学専攻専攻長．2015 年 JST CREST 研究代表者．2014 年日本ソフトウェア科学会基礎研究賞．2014 年日本学術振興会賞受賞．2013 年文部科学大臣表彰科学技術賞受賞．2007 年文部科学大臣表彰若手科学者賞受賞．情報処理学会長尾真記念特別賞受賞．2006 年 AAMAS2006 最優秀論文賞受賞．2005 年日本ソフトウェア科学会論文賞受賞．平成 16 年度 IPA 未踏ソフトウェア創造事業スーパークリエータ認定．AAMAS2013 Program Chair．マルチエージェントシステム国際財団（IFAAMAS）理事．

アーメッド ムスタファ

ウーロンゴン大学大学院博士課程修了．アデレード大学客員研究員，オークランド工科大学客員研究員，および Data61 客員研究員を経て，現在，名古屋工業大学准教授．自動交渉，マルチエージェント強化学習，マルチエージェント社会における信頼と評判，深層強化学習，サービス指向コンピューティング，集合知，知能交通システム，データマイニングに興味がある．ICWS，ICSOC，および WWW 運営審査委員の担当経験がある．人工知能学会，IEEE Computer Society，Australia Computer Society，および Service Science Society of Australia 各会員．