

人狼ゲームエージェントにおける行動選択手法の比較

王 天鶴^{1,a)} 金子 知適^{2,3,b)}

概要: 人狼ゲームとは、プレイヤー間の情報が非対称という特徴がある不完全情報ゲームの一種である。近年、人工知能は完全情報ゲーム分野で良い成果をあげた一方で、不完全情報ゲームについての研究は発展の余地がある。本論文では、強化学習の一種である Q 学習を人狼ゲームのエージェントに適用するためのモデルを設計し、 ϵ -グリーディ法、Softmax 法、UCB 法、一様ランダム法を行動選択手法とし、村人、占い師、人狼のエージェントを実装した。各行動選択手法ごとに、3 - 4 人と変えながら学習したエージェントの勝率を評価し、Q 学習モデルと行動選択手法の効用を比較、検討した。それらの greedy 法での勝率を評価した結果、3 人ゲームで、学習段階で ϵ -グリーディ法を利用したエージェントは学習後の勝率が一番高かった。4 人ゲームでは、学習段階で一様ランダム法を利用した村人エージェント、 ϵ -グリーディ法を利用した人狼エージェント、UCB 法を利用した占い師エージェントの学習後の勝率がそれぞれの役職において勝率が一番高かった。

Comparison of Methods for Choosing Actions in Werewolf Game Agents

TIANHE WANG^{1,a)} TOMOYUKI KANEKO^{2,3,b)}

Abstract: Werewolf, also known as Mafia, is a kind of game with imperfect information that features information asymmetry. Recently, artificial intelligence achieved excellent performance in the world of perfect-information games, however, there are still room for development of imperfect-information games. In this paper, we implemented the agents of Villager, Seer and Werewolf in which we applied our model on the basis of Q-learning, a reinforcement-learning technique, and compared ϵ -greedy, Softmax, UCB and Uniform Random methods for choosing actions. For each method, we evaluated it in 3-to-4-player games by the win ratio of agents after learning multiple cases. The experimental results by evaluating win ratio in greedy method showed that the agent learned in ϵ -greedy performed best in 3-player games while in 4-player games, the Villager agent used Uniform Random method, the Werewolf agent used ϵ -greedy method and the Seer agent used UCB method in stage of learning performed best after learning respectively.

1. はじめに

近年、完全情報ゲームにおける人工知能の発展は顕著であり、オセロやチェスに続き、囲碁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 [2]。一方で、観測できない情報がある前提でプレイヤーが行動する不完全情報ゲームの人工知能についての研究は発展の余地があり、新たな研究方向の一つである。

人狼ゲームは、書籍 [6] を紹介すると、プレイヤーが善き村人とそれを狙う人狼に分かれ、村人は人狼を排除することで、人狼は村人を襲い村から排除することで勝利を目指すというゲームである。ただし、村人となったプレイヤーはどのプレイヤーが人狼なのかを知らない。そのため、疑いながら人狼らしきプレイヤーを排除していく。一方で、人狼となったプレイヤーたちはお互いが人狼であることを

¹ 東京大学大学院学際情報学府
Graduate School of Interdisciplinary Information Studies,
The University of Tokyo

² 東京大学大学院情報学環
Interfaculty Initiative in Information Studies, The University of Tokyo

³ 国立研究開発法人科学技術振興機構 さきがけ
JST, PRESTO

^{a)} wangtianhe@g.ecc.u-tokyo.ac.jp

^{b)} kaneko@acm.org

認識したうえで、自分たちが人狼であることを隠し、村人の振りをしながら村人であるプレイヤーを襲撃していくというゲームである。

これまでに、完全情報ゲームにおける人工知能の研究が大きく進歩し、ゲーム木探索を応用するエージェントは、人間のトップレベル以上の成績に到達した。一方で、不完全情報ゲームは、プレイヤーが局面の状況を完全に把握することができないので、既存のゲーム木探索手法をそのまま適用することはできない。

本研究では、行動とプレイヤーの数が多く、複雑性の高い人狼ゲームを対象にエージェントの戦略を研究する。目的は各種の学習手法を応用し、それらの性能を評価、比較し、得られた結果を総合して強いエージェントを作成することである。UCBQ アルゴリズム [4] を適用し、従来手法と比較しながら、検討していく。

2. 関連研究

研究 [3] ではプラットフォーム [5] において、 ϵ -グリーディ法を行動選択方法とした Q 学習で、学習エージェントを設計した。

その研究が提案した人狼での Q 学習手法は、CO 状況、占い結果、霊媒結果、投票結果から得た情報に基づいて、各プレイヤーの可能な役職のパターンを全て算出し、それらのいずれかを状態として扱う。CO 条件の選択、人狼陣営の騙り役職の選択、投票などの対象選択を行動として扱う。学習の結果として、ランダムプレイヤーと比較し、全ての役職について、その役職側の勝率が上昇している。

人狼ゲーム以外の研究で、研究 [4] では Q 学習に UCB アルゴリズムを組み合わせた UCBQ というアルゴリズムが提案されている。その性能は、連続空間における非マルコフ性経路探索問題において、従来手法と比べ良く機能することが示されている。

ある状態の UCB 値は下記の式で表される [1]。

$$UCB_i = \bar{X}_i + C \sqrt{\frac{\ln n}{n_i}} \quad (1)$$

i は行動の番号、 n はその状態での今までの行動回数、 n_i は行動 i が選択された回数を意味する ($\sum_i n_i = n$)。人狼ゲームの場合、プレイヤーの占う、攻撃と、投票することを行動とする。また、 $\bar{X}_i$ は行動 i のその時点までの経験から推定した未来の報酬の予測値であり、 C は exploration と exploitation のバランスをとるための定数である。

UCB 値が高い行動を試行することで、試した回数が少ないものの、報酬値が高い可能性がある行動を、探索することができる。

以下には文献 [4] の内容を、表現を一部変更し説明する。UCBQ アルゴリズムでは、UCB 値の第一項を Q 値に置き換え、Q 学習の試行において行動を選択する。その際に、 ϵ の確率でランダムに行動を選択し、 $1 - \epsilon$ の確率で、過去

に選択されたことのある行動の中から最大の UCB 値を持つ行動を選択するという手法をとる。

3. 提案手法

本研究では研究 [3] のエージェントの学習手法を参考にし、文献 [5] で構築された人狼ゲームのプラットフォームにおいて、文献 [4] の UCB 法を行動選択法として Q 学習を行うエージェントを提案する。そのエージェントを、研究 [3] と同様、 ϵ -グリーディ法及び softmax 法を行動選択法とするエージェントを評価して比較する。また、Q 学習手法での状態の扱いと行動の設計を提案する。

提案した Q 学習エージェントの状態、行動、報酬等の設計について説明する。

3.1 状態

実際の人狼ゲームの局面において、会話の内容は各プレイヤーにとって重要な情報源である。自己身分の主張、他のプレイヤーへの疑いと身分推測が会話の主要部分である。プレイヤーは会話の内容に基づいて投票、攻撃などの対象を選ぶことと仮定する。

この仮定に従い、プレイヤーたちのお互いの態度、CO 状況、検証した事実を状態内容として提案する。

プレイヤーたちのお互いへの態度は、あるプレイヤーの占い結果、推測、投票予告に関する発話を抽出し、他の各プレイヤーがどの陣営の人と想定しているかに対応する。例えば、プレイヤー A が「プレイヤー B が人狼であると思う」と発話し、プレイヤー B がプレイヤー C を占い、「C は人間陣営である」と発話した場合、プレイヤー A はプレイヤー B への態度が人狼陣営、プレイヤー B はプレイヤー C への態度が人間陣営という態度とそれぞれみなす。

CO 状況は、各プレイヤーのカミングアウト発話から抽出する。自分のカミングアウト役職も CO 状況として扱うが、人狼の場合は騙り役職を扱う。

確定した事実は占い師、人狼のみ入手可能な情報である。占い師は、毎晩の占い結果を確定した事実として得る。人狼は他の同じく人狼であるプレイヤーの身分を、確定した事実として持つ。

表 1 のように、プレイヤー数が N の場合、上記 3 種類の情報を $N \times N$ の行列で格納する。右下向きの対角線にある要素は CO 状況である。自分のインデックスを i とし、他のプレイヤーのインデックスを j とし、 i 行目 j 列目にある要素は確定した事実である。その他の要素はプレイヤーたちのお互いへの態度である。

3.2 行動

プレイヤーは会話から局面の状況を推定し、自分の戦略で投票及び占いと攻撃の対象を定め、対象の選び方を Q 学習で学ぶことを提案する。本研究は下記の 6 種類の選び方

表 1 状態を格納する行列 (プレイヤー 0 が占い師の場合)

	プレイヤー 0	プレイヤー 1	プレイヤー 2
プレイヤー 0	CO 状況	1 に関する事実	2 に関する事実
プレイヤー 1	1 の 0 への態度	CO 状況	1 の 2 への態度
プレイヤー 2	2 の 0 への態度	2 の 1 への態度	CO 状況

を用意した。

(1) 疑う人

ある状態において、他のプレイヤーを人狼陣営と推測する発話が一番多い人にする。

(2) 疑われる人

ある状態において、他のプレイヤーに人狼陣営と推測された発話が一番多い人にする。

(3) 矛盾の相手

あるプレイヤーへの態度の分岐が一番多い二人の中に、自分以外の任意の一人にする。

(4) 占い信用無し

占い師、霊能者 (8 人以上のプレイヤーは必要なので、本研究では扱わない) としてカミングアウトしたプレイヤーから任意の一人にする。

(5) 無言な人

発話が一番少ない人にする。

(6) ランダム

生きているプレイヤーから任意の一人にする。

上記の選び方は、状態によっては、人間陣営か人狼陣営かによって有利な行動と不利な行動が両方存在する。

3.3 報酬と Q 値の更新

研究 [3] の手法を参考し、報酬値はゲーム終了する際に与える。自分の陣営が勝利した場合に 100 を与え、敗北した場合に 0 を与える。

ゲーム中は、毎回の行動選択時の状態と行動を記録し、ゲーム終了した際に報酬値を得て、ゲーム中に選んだ行動の Q 値を更新する。

3.4 CO 条件の設定

人間陣営の村人と占い師は他のプレイヤーに人狼と疑われると、自分の役職を CO すると設定する。

人狼陣営の人狼は他のプレイヤーに人狼と疑われると騙り役職として CO する。騙り役職はエージェント初期化する際に人間陣営の任意に指定する。

4. 実験考察

本研究では文献 [3] の実験を参考し、文献 [5] で構築された人狼ゲームのプラットフォームにおいて、提案手法の学習と評価の実験を行う。

4.1 学習段階

今回の実験は 3-4 人の人狼ゲームと変えながら行う。プラットフォーム [5] のデフォルト割当に従い、占い師、人狼、村人 3 種類の役職があり、4 人の場合に村人は 2 人いる。Q 学習における状態のサイズはプレイヤー人数と関係あるため、プレイヤー人数ごとに学習を別々に行う。エージェント以外の全てのエージェントはプラットフォームが用意した Sample Player である。

5 種類の人数設定において、各役職のエージェントに対し、 ϵ -グリーディ法、Softmax 法、UCB 法、一様ランダムを行動選択法とするエージェントを作成した。そこで、エージェント毎に 100000 回の試行を行う。

試行回数が増えながら、勝率の変化を記録し、下記のグラフのように示す。(各グラフによって、それぞれの縦軸範囲がある。)

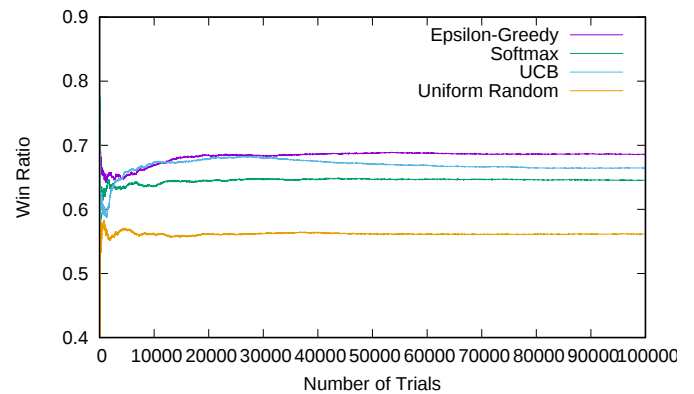


図 1 村人勝率の学習試行回数による変化 (3 人ゲーム)

図 1 に示すように、4 個の村人エージェントは 3 人ゲームにおいて、学習回数が 20000 回前後で勝率が収束した。最初に ϵ -グリーディ法と UCB 法のエージェントの勝率はほぼ等しいが、学習回数が 100000 万になるまで UCB 法の勝率が下がり、差が段々大きくなる。勝率の順位は良い方から ϵ -グリーディ法、UCB 法、Softmax 法、一様ランダムである。学習完了時に ϵ -グリーディ法の勝率は一様ランダムよりおよそ 12 ポイント高いことが分かった。

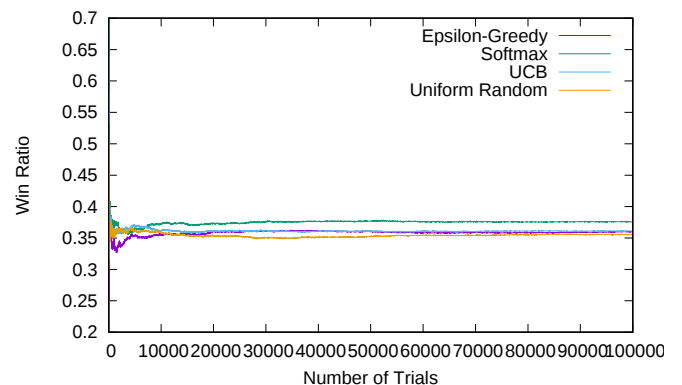


図 2 人狼勝率の学習試行回数による変化 (3 人ゲーム)

図 2 に示すように、4 個の人狼エージェントは 3 人ゲームにおいて、学習回数が 10000 回前後で勝率が収束した。勝率の順位は良い方から Softmax 法, UCB 法, ϵ -グリーディ法, 一様ランダムである。学習完了時に Softmax 法の勝率は一様ランダムよりおよそ 2 ポイント高いことが分かった。

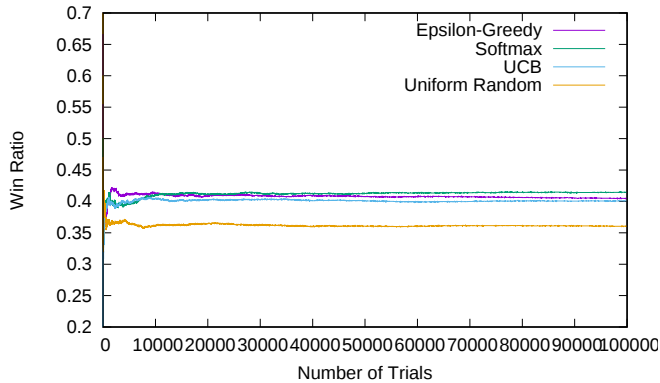


図 3 占い師勝率の学習試行回数による変化 (3 人ゲーム)

図 3 の示すように、4 個の占い師エージェントは 3 人ゲームにおいて、学習回数が 10000 回前後で勝率が収束した。勝率の順位は良い方から Softmax 法, ϵ -グリーディ法, UCB 法, 一様ランダムである。学習完了時に Softmax 法の勝率は一様ランダムよりおよそ 5 ポイント高いことが分かった。

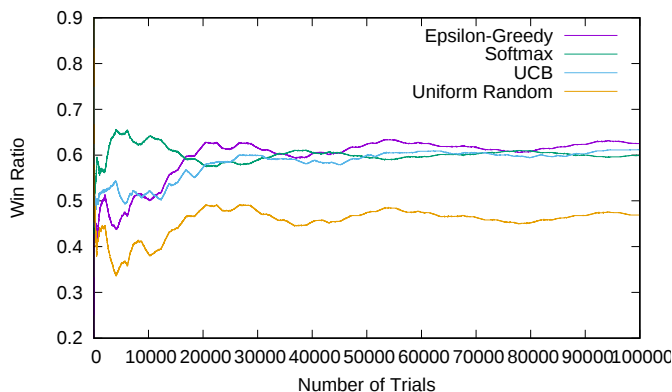


図 4 村人勝率の学習試行回数による変化 (4 人ゲーム)

図 4 に示すように、4 個の村人エージェントは 4 人ゲームにおいて、勝率が全体的に 3 人ゲームより激しく変化してきた。学習回数が 60000 回前後で勝率が収束した。勝率の順位は良い方から ϵ -グリーディ法, UCB 法, Softmax 法, 一様ランダムである。学習完了時に Softmax 法の勝率は一様ランダムよりおよそ 16 ポイント高いことが分かった。

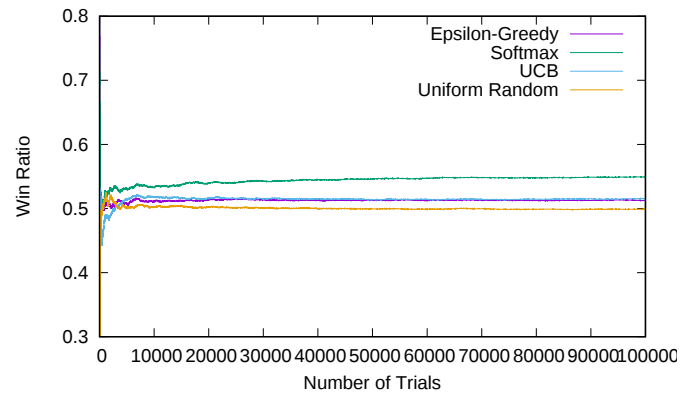


図 5 人狼勝率の学習試行回数による変化 (4 人ゲーム)

図 5 に示すように、4 個の人狼エージェントは 4 人ゲームにおいて、学習回数が 6000 回前後で勝率が収束した。勝率の順位は良い方から Softmax 法, UCB 法, ϵ -グリーディ法, 一様ランダムである。学習完了時に Softmax 法の勝率は一様ランダムよりおよそ 5 ポイント高いことが分かった。

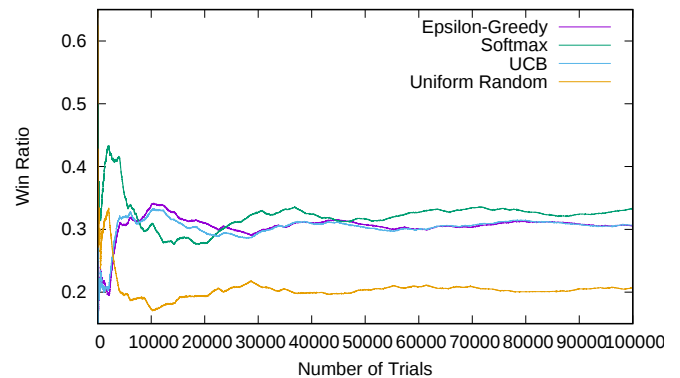


図 6 占い師勝率の学習試行回数による変化 (4 人ゲーム)

図 6 に示すように、4 個の占い師エージェントも 4 人ゲームにおいて、勝率が 3 人ゲームより激しく変化してきた。学習回数が 50000 回前後で勝率が収束した。勝率の順位は良い方から Softmax 法, UCB 法, ϵ -グリーディ法, 一様ランダムである。学習完了時に Softmax 法の勝率は一様ランダムよりおよそ 13 ポイント高いことが分かった。

各エージェントにおいて、Softmax 法, UCB 法, ϵ -グリーディ法は全体的に一様ランダムより勝率が高いと観察した。ただ、占い師と村人のエージェントにおいて一様ランダムとの差が人狼のエージェントより顕著である。

また、各行動選択法を比べれば、占い師と人狼のエージェントは Softmax 法を応用した場合の勝率が最も高く、村人のエージェントは ϵ -グリーディ法を応用した場合の勝率が最も高いと観察した。UCB 法を応用した場合の勝率は ϵ -グリーディ法と Softmax 法の間にある。

エージェントはゲーム進行中に、一言の発話が増えても状態の変化であるので状態を更新するが、エージェントが

保存する状態は行動を選択する時に現れた状態のみである。学習が終了した際に、各エージェントが保存している状態数を表 2 と表 3 のように示す。

表 2 学習終了時の状態数 (3 人ゲーム)

	村人	人狼	占い師
ϵ -グリーディ法	44	15	110
Softmax 法	44	15	103
UCB 法	44	15	111
一様ランダム法	44	15	124

3 人ゲームの場合、各エージェントの可能な状態数の理論値について、村人は 2592 個、人狼は 3888 個、占い師は 18144 個である。

表 3 学習終了時の状態数 (4 人ゲーム)

	村人	人狼	占い師
ϵ -グリーディ法	227	202	1700
Softmax 法	227	202	1631
UCB 法	227	202	1610
一様ランダム法	227	202	1802

4 人ゲームの場合、各エージェントの可能な状態数の理論値について、村人は 2519424 個、人狼は 3779136 個、占い師は 37791360 個である。

学習完了時にエージェントが保存している状態数が理論値より極めて小さいという現象の原因は二つあると考える。

1 つ目は、学習エージェントが Sample Player と対戦し、Sample Player の行動は有限であり、基本的に現れない行動 (例えば、人狼の身分で CO する) も理論値に算入したためである。

2 つ目は、学習エージェントが保存している状態は、行動を選択する際 (投票, 占う際) の状態のみである。プレイヤーが一言の発話をしても、状態を更新するが、保存はしない。発話過程に現れた状態は保存した状態より数が多いので、理論値に参入した状態の大部が保存されていないためである。

4.2 評価段階

学習したエージェントの Q 値と UCB 値を利用し、エージェントの評価を行う。 ϵ -グリーディ法, Softmax 法, 一様ランダム法を利用して学習したエージェントは Q 値が一番大きい行動を選択する。UCB 法を利用したエージェントは UCB 値が一番大きい行動を選択する。

各エージェントで 10000 回のゲームを試行し、勝率の統計は表 4 と表 5 のように示す。

表 4 学習後の勝率評価 (3 人ゲーム)

	村人	人狼	占い師
ϵ -グリーディ法	70.0%	37.2%	39.6%
Softmax 法	60.8%	37.1%	36.0%
UCB 法	66.7%	35.8%	38.7%
一様ランダム法	66.4%	36.0%	39.6%

ϵ -グリーディ法を利用したエージェントの勝率評価は他の行動選択手法より高いと観察した。ただ、占い師の場合、 ϵ -グリーディ法と一様ランダム法を利用したエージェントの勝率は等しいと評価した。

表 5 学習後の勝率評価 (4 人ゲーム)

	村人	人狼	占い師
ϵ -グリーディ法	66.8%	52.7%	28.1%
Softmax 法	57.3%	50.1%	21.3%
UCB 法	53.8%	51.8%	52.2%
一様ランダム法	70.3%	52.2%	31.2%

一様ランダム法を利用した村人エージェント, ϵ -グリーディ法を利用した人狼エージェント, UCB 法を利用した占い師エージェントがそれぞれの役職において勝率が一番高いと評価した。

学習段階の勝率と比較し、一様ランダム法の勝率評価が良くなる現象の原因は、評価段階でエージェントの行動選択手法を「Q 値が一番高い行動を選択する」という手法に変更したためである。この手法に変更すると、学習段階で生成した Q 値の精度が高いならば、評価段階で勝率の評価も高くなる。このため、各行動を等しく確率で選択する一様ランダム法を利用して、エージェントが生成した Q 値の精度も高いと考える。

5. 結論

本研究は、人狼ゲームで新たな状態行動モデルで UCB 法を行動選択手法とする Q 学習のエージェントを提案した。そこで、 ϵ -グリーディ法, Softmax 法及び一様ランダム法を対照しながらエージェントの実験と評価を行った。

その結果、3 人ゲームの場合、 ϵ -グリーディ法を利用したエージェントは学習後の勝率が一番高かった。4 人ゲームの場合に、一様ランダム法を利用した村人エージェント, ϵ -グリーディ法を利用した人狼エージェント, UCB 法を利用した占い師エージェントの学習後の勝率がそれぞれの役職において勝率が一番高かった。

しかしながら、現段階設計した状態は人狼の局面情報の 3 つの側面だけを格納するし、行動の種類も豊富になる余地がある。また、人数が増えると状態のサイズが大きくなるため、計算効率の向上もやや難しくなる。今後の課題としては、局面情報を更に格納する状態と豊富な行動を設計する。それに、計算効率を多い人数に対応する学習アルゴリズムを検討する。

謝辞

この研究の一部は、JSPS 科研費 16H02927 と JST さきがけの支援を受けています。

参考文献

- [1] Sutton, R. S. and Barto, A. G.: *Reinforcement learning: An introduction*, Vol. 1, No. 1, MIT press Cambridge (1998).
- [2] Technologies, D.: DeepMind, <https://deepmind.com/>. Accessed July 24, 2017.
- [3] 梶原健吾, 鳥海不二夫, 稲葉通将, 東京大学: 人狼における強化学習を用いたエージェントの設計, 人工知能学会全国大会論文集, Vol. 29, pp. 1–3 (2015).
- [4] 齊藤晃貴, 野津亮, 本多克宏: 強化学習における UCB 行動選択手法の効果, 日本知能情報ファジィ学会 ファジィシステム シンポジウム講演論文集 第 30 回ファジィシステムシンポジウム, 日本知能情報ファジィ学会, pp. 174–179 (2014).
- [5] 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, Vol. 2015 (2015).
- [6] 鳥海不二夫, 片上大輔, 大澤博隆, 稲葉通将, 篠田孝祐, 狩野芳伸: 人狼知能だます・見破る・説得する人工知能 (2016).