強化学習を活用したマンカラの最善手の探索

東京都立立川高等学校 2年 山本 勇太

1. マンカラとは

マンカラは世界最古のゲームの一つとして世界中に100種類以上ものルールが存在するともいわれているボードゲームの一種。代表的なルールとして、世界各地で遊ばれているKalah、日本のBasic、フィリピンのSungkaなどがある。

2. 研究背景

マンカラで勝てたとき、その勝因が分からないことが多いと感じた。そのため、どのようにしたら勝てるのかを調べようと思った。先行研究から、特定条件下では必勝戦略が考えられているが、石の数が増えると条件に当てはまらない場合が指数関数的に増加するなどの課題があることが分かった。また、勝利時の盤面から逆算して必勝戦略が使える盤面を生成するといったことも行われていたが、同様に条件の制約が不可欠であった。

3. 目的

今回の研究では、マンカラの強化学習を行い、その打ち方から特徴を見出すことで、盤面の数が指数サイズになっても初期盤面からの最善手を特定できるようにすることを目的とする。

4. 今回のルール

ルール[A]は次の条件のうち、
3以外の条件に従い、
ルール[B]は次のすべての条件に従う。

1. 図1のように盤面の各ポケットに3つずつ石を入れる。
   ※ポケットは各プレイヤー3つ、向かって右側にストア1つ。
2. プレイヤーは交互にポケットを選び、その中の石を反時計回りに1つずつ入れていく。(相手のストアには入れない)
3. [B]のみ。条件2において、最後の石が自分のストアに入ったらもう一度自分の番になる。
4. 先に自分のポケットの石を全てなくしたプレイヤーの勝ち。

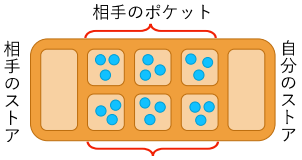

図1　盤面

5. 強化学習

[A],[B]それぞれのルールの下で、強化学習を行うプレイヤーとランダムにポケットを選ぶプレイヤーを10万回対戦させた。これを10セット行い、各セットでの強化学習の勝率の推移の平均を求めた。強化学習にはQ学習(ε-greedy法)を用い、対戦やグラフの作成、5.分析 や 6.検証 は、Pythonを用いて行った。

● ルール[A]

10万回の対戦終了時点でQ学習の勝率推移の平均が約75%に達した(図2)。乱雑度0の状態での対戦においては約95%となった。

● ルール[B]

10万回の対戦終了時点でQ学習の勝率推移の平均が約70%に達した(図3)。乱雑度0の状態での対戦においては約93%となった。

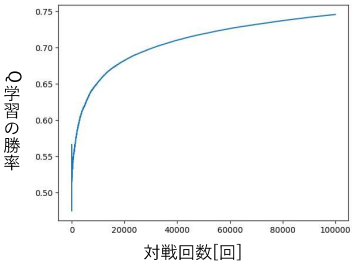

図2　Q学習の勝率推移の平均[A]

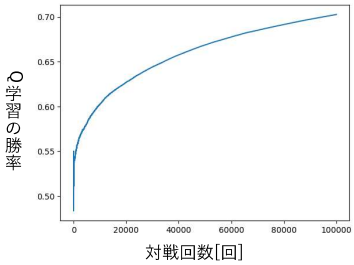

図3　Q学習の勝率推移の平均[B]

※乱雑度
Q学習(ε-greedy法)では、学習中にそれまでの行動とその結果をもとに、より良い行動を選択するように行動していく。この時、一度うまくいった行動を永遠に繰り返してしまう可能性があるので、それを防ぐために一定の割合(=乱雑度)でランダムに行動するように設定する。今回は乱雑度0.3に設定した。

6. 分析

強化学習の際、盤面の状況やQ学習の各行動(各ポケット)に対する報酬とQ値、直前の相手の行動(ルール[A]のみ)を表の形で記録した。この表をもとに以下の2つの分析を行った。

※報酬…ポケットを選んだ結果、勝利したときは100点、敗北したときは-100点というように得点を与え、Q学習では報酬の和が最大となるように(≒最も高いQ値をもつ)ポケットを選ぶ。
※Q値…報酬から計算される各行動(各ポケット)の価値であり、値が大きいほど勝利に近づきやすいと判断できる。

●分析①

ルール[A],[B]ともに、一番最初の手における各ポケットのQ値を比較した(表1)。
ルール[A],[B]ともに、左とそれ以外のポケットのQ値に差が見られるため「左のポケットを選ぶほうが他の2つよりも勝ちやすい」と考えられる。

表1　初手のQ値

	左	中央	右
[A]	87.9	71.9	72.5
[B]	64.1	41.6	54.7

●分析②

ルール[A]において、直前の相手の打った手ごとに平均した各ポケットのQ値平均を比較した(表2)。
直前の相手の打った手によってQ値の大きいポケットが異なるため「直前の相手の打った手に応じて自分の打つ手を変えたら勝ちやすい」と考えられる。

表2　相手の手ごとに平均したQ値

	左	中央	右
左(相手)	7.2	9.5	10.5
中央(相手)	1.9	3.3	8.9
右(相手)	12.7	10.3	10.2

7. 検証

分析結果から得られた考えが正しいか検証するために以下の2つの戦略で打つプレイヤーとランダムに打つプレイヤーをルール[A]の下で対戦させ、その勝率を求めた。

●“左”戦略

分析①より、選べるポケットのうち、一番左のものを選択する戦略。
1万回の対戦終了時点で“左”戦略の勝率は約52%となった(図4)。

●“相手”戦略

分析②より、直前の相手の手の打った手に応じて、表2のQ値が大きいポケットを選択する戦略。
5000回の対戦終了時点で“相手”戦略の勝率は約53%となった(図5)。

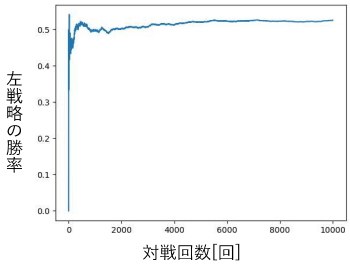

図4　“左”戦略の勝率推移

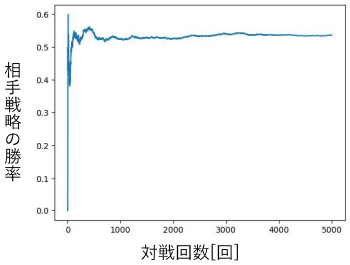

図5　“相手”戦略の勝率推移

8. 考察

強化学習の分析や検証を総合して考察すると、Q値の違いから「一番左のポケットを選ぶ」「相手の手に応じてポケットを選ぶ」といった戦略が考えられるが、このような戦略のみでは勝率を上げることにはつながらないため、序盤・中盤・終盤に合わせた戦略など、より複雑な戦略が必要だと考えられる。

9. 展望

今後は、序盤・中盤・終盤に分けた分析やそれに合わせた戦略の検証など、強化学習の打った手の分析と検証を継続する。加えて、マンカラのルールを変えて同様の分析・検証を行い、ルールと勝敗の関係がないか調べる。また、強化学習の各種設定を変えたり、学習の対戦相手としてランダムに選ぶプレイヤーだけでなく強化学習をするプレイヤーや人間を用意したりすることによってより強いプレイヤーを作成し、最善手の探索に活用する。

10.まとめ

マンカラの盤面の数が指数サイズになっても初期盤面からの最善手が特定できるよう、Q学習の打った手を分析することで最善手を見つけることを試みた。Q学習を行った結果、勝率は9割以上に達し、打った手の分析から左のポケットの選択や相手に応じた選択が有効だと考えられた。しかし、これらの戦略では勝率が上がらなかったことから、盤面・戦況に応じたより複雑な分析や複雑な戦略の考案が必要だと考えられる。

11. 参考文献

- マンカラルール - 公益財団法人日本レクリエーション協会 , https://recreation.or.jp/activities/genki_up/mancala/.
- 誰かに言いたい 世界のアソビ裏話 | Nintendo Switch - 任天堂 , https://www.nintendo.co.jp/switch/as7ta/discovery/second/article01.html.
- マンカラ開発記 | S.K | note , https://note.com/sk_game_theory/m/m97f73b55a693.
- 伊藤真 『ScratchでAIを学ぼう ゲームプログラミングで強化学習を体験』 , 日経BP , 2020.
- 前井康秀, 木谷裕紀, 土中哲秀, 小野廣隆 「Simple-Kalahにおける勝敗確定の十分条件」, 火の国情報シンポジウム2018論文集, A3-2, 2018.
- 小川遼, 小泉康一, 大規正伸 「二人零和有限確定完全情報ゲームの先手勝利盤面生成手法に関する研究」, 情報処理学会研究報告, Vol.2019-GI-41, No.23, pp.1-7, 2019.