大貧民における方策・価値評価関数の学習とその利用法

福澤　詩音　　【 高度プログラミング研究室 】

1　はじめに

多人数・不完全情報ゲーム「大貧民」は，ランダム性が強いことから，単純なモンテカルロ木探索では有効な手を見つけることが難しい．本研究では，大貧民における方策・価値評価関数を学習し，その情報を用いて有効な手を見つける手法を提案する．また，不完全情報ゲームにおける強化学習の有効性についても検証する．

2　ネットワーク構成

ネットワークは，方策ネットワークと価値評価ネットワークの 2 つから構成される．方策ネットワークは，現在の局面における各手の選択確率を出力する．価値評価ネットワークは，現在の盤面における提出可能な手のそれぞれについて，その手を選択した場合のスコアを予測する．

ニューラルネットワークへの入力は，以下の要素で構成される．

場の状態: 縛り，革命，場に出ているカードのスート
手札の情報: 自分の手札，全プレイヤの使用済みカード
場・提出カード: 場のカード，提出可能なカード
プレイヤ情報: 全プレイヤの手札枚数，ランク

全プレイヤの使用済みカード，手札枚数，ランクに関して，自分を先頭とした順番に情報を並べる．

ネットワークの構成は全層結合で，活性化関数には ReLU を用いる．方策ネットワークの出力層には Softmax 関数を用い，各手の選択確率を出力する．価値評価ネットワークの出力層には線形関数を用い，各手のスコア予測値を出力する．

3　実験

モデルの学習は，教師あり学習と強化学習を組み合わせて行う．まず，ルールベースプレイヤ 5 体の対戦データを用いて教師あり学習による初期モデルを構築する．その次に，初期モデルを用いて自己対戦を行い，強化学習によってモデルを強化する．

教師あり学習は，ルールベースプレイヤ 5 体の 100000 試合分の対戦データを用いて行う．強化学習は，ルールベースプレイヤ 2 体，初期モデル 2 体，強化学習クライアント 1 体を用いて 600000 試合の対戦を行う．

提出手の選択は，方策ネットワークから出力された各候補手の選択確率に対して対数確率を取ったものと，価値評価ネットワークから出力された値の合計比率が 1 になるようにした値を用いる．

図 1 は，方策と価値評価の比率をそれぞれ変化させたものに対して，ルールベースプレイヤ 4 体と対戦した時の平均スコアである．図の表 1 は，学習回数に対する平均スコアの推移である．

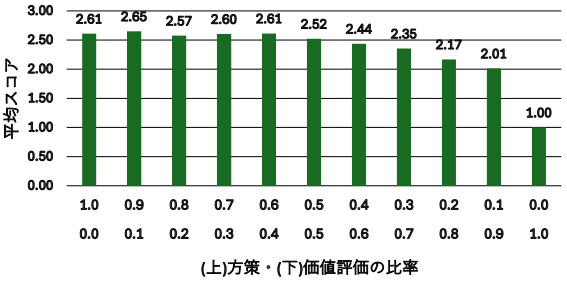

図 1　各比率における平均スコア

学習回数	0	250	500	750	1000
平均スコア	3.04	3.02	3.02	2.96	

表 1　学習回数と平均スコアの推移

4　まとめ

本研究では，大貧民における方策・価値評価関数を学習し，その情報を用いて有効な手を見つける手法を提案した．実験の結果，方策・価値評価関数の両方を利用することで，ルールベースプレイヤに対して優位に立つことはできなかったものの，それぞれの組み合わせがスコアに影響を与えることが確認できた．また，強化学習によるクライアントの強化も見られた．今後の課題として，強化学習の手法の改善や，ネットワーク構成の改良，対戦回数の増加などが挙げられる．

参考文献

[1] 内田 順平，穴田 一，“ニューラルネットワークによる大貧民のカード提出モデル構築”，第 36 回ファジィシステムシンポジウム 講演論文集

[2] 電気通信大学，2020 年，“UECda コンピュータ大貧民大会”，https://flute.u-shizuoka-ken.ac.jp/daihinmin/2023/document_rules.html，2026 年 2 月閲覧．