証明数・反証明数を用いた
罰回避政策形成アルゴリズムの高速化に関する研究

登 崇志 (310083)

北陸先端科学技術大学院大学 情報科学研究科

2005 年 2 月 10 日

キーワード: 強化学習, 罰回避政策形成アルゴリズム, 証明数・反証数, ニューラルネットワーク, オセロ.

強化学習とは，報酬という特別の入力を手がかりに環境に適応する機械学習システムである．明示的に正解を与える教師なしに学習できるので魅力的な枠組みと言える．つまり，設計者が何をすべきか (what) をエージェントに報酬という形で指示しておけば，どのように実現するか (how to) をエージェントが学習によって自動的に獲得する枠組となっている．

ゲーム問題において，方策 (現状態に対して行動を決定する関数) を考案するには，そのゲームに対する十分な知識が必要である．しかし，強化学習を用いることにより，勝敗を報酬とすることで自動的に方策を獲得することができる．強化学習の適用例として，オセロゲームの盤を表現したニューラルネットワークを，TD(λ) で学習するもの，チェスプログラム Knight Cap の線形評価関数の各重みを，TD(λ) で調整するものなどがある．これらは最適政策すなわち，最善手を獲得することを目的としている．

一般的な 2 人対戦のゲーム問題は負けない事，すなわち罰の回避が大前提とされ，必ずしも圧勝等の最適性は要求されない．このような特性を利用したゲーム問題の強化学習の適用例として，宮崎による罰回避政策形成アルゴリズムを用いたオセロゲームへの応用があげられる．罰回避政策形成アルゴリズムとは，それまでに経験した罰 (負け, 負の報酬) を得る状態と行動を記憶して，それを回避する政策を形成するアルゴリズムである．ある状態で選択可能な全ての行動が罰ルールとなったときに，その状態を罰状態とする．また，相手が最善の行動を行ったときに罰状態へ至ると確定した行動も罰ルールとして，罰ルールを伝播していく．しかし，膨大な状態空間でこれを応用する場合，各状態の罰の伝播にかかる計算コストが膨大となる．また，勝敗が確定するまで罰を得ることが無い問題では，序盤から中盤にかけて罰の伝播が行われずに試行回数が膨大となるなどの問題点がある．そのため，これらの問題を解消するために以下の改良がなされた．現在経験している状態と，現在の状態から数手先を読んで展開した状態だけから罰を伝播し，罰の伝播

にかかる計算コストを軽減した．また，既存知識を与えることによりヒューリスティックな罰 (準罰) を与え，序盤と中盤において罰の伝播を可能にした．しかし，準罰は必ずしも正確な罰の規範ではないため，以下のような問題が生じる．罰の伝播を準罰に頼っている序盤から中盤において，準罰が少ないときは，勝ち筋を獲得するまでに必要な罰の伝播に遅延が起こる．逆に多いときは，誤った伝播が頻繁におこり，開始状態までが準罰となり勝ち筋の獲得は計算コストの問題でできない結果となった．つまり，罰回避政策形成アルゴリズムによる勝ち筋の獲得には，ヒューリスティックによる罰の精度に大きく依存してしまうのである．

本稿では宮崎による罰回避政策形成アルゴリズムにおいて問題となる，ヒューリスティックの評価に大きく依存した試行回数の軽減を目的とし，先読みと罰の伝播に証明数と反証数を用いた手法を提案する．(反) 証明数とは，その状態，つまり局面を勝ち (負け) を証明するために必要な最終局面 (勝敗が決定した状態) の数である．証明数・反証数をオセロゲームに適用すると，(反) 証明数はその盤面が勝ちであるために勝ち (負け) を示さないといけない先端盤面の個数の最小値となる．先端の (反) 証明数の個数が最小の手は現在の状態を (負け) 勝ちと示すために，必要な状態展開数が最も少ない手となる．つまり，展開する状態が少ないため，深く探索する事ができ終了盤面により近い状態を先読みする事が可能となる．この手法は勝敗が決する終盤の探索問題 (勝敗の探索) に使われてきた手法である．罰回避政策形成アルゴリズムは最終状態の罰 (負け) を伝播していくアルゴリズムであるので，この手法をもちいて罰の伝播を効率することができる．

提案手法においては，伝播すべき罰を最終局面とみなし，証明数と反証数を罰の伝播に用いることにより，現在の状態が罰 (必負) であるか高速に判断しする．それにより，学習速度が向上し，少ない試行回数で罰回避政策の獲得が可能となる．その結果，試行回数の削減により，ヒューリスティックな評価による罰が少ない，罰の伝播に試行回数が多くかかる問題においても，ヒューリスティックな評価にあまり頼らずに罰回避政策を獲得することができる．

本研究では，証明数と反証数を用いた罰回避政策アルゴリズムを実際にオセロゲームに実装した．その結果，提案手法が従来手法に比べ少ない試行回数で罰回避政策 (勝ち筋) を獲得することができた．また，従来の手法ではヒューリスティックな評価による罰を少ないために，計算コストの問題で罰回避政策 (勝ち筋) を見つけることをできなかった問題に大しても，提案手法では獲得することに成功した．

本研究ではさらに，学習により獲得した政策の汎化を行った．罰回避政策形成アルゴリズムは特定の相手プレイヤーに対して勝ち筋を発見する方法である．よって，学習結果を類似問題といえど反映させることはできない．そこで，異なる対戦相手に対して，過去の類似問題での学習成果を汎用するために，ニューラルネットワークの評価によって罰を判定する手法を提案する．盤面の罰判定をニューラルネットワークで表現し，これに強化学習によって得られる罰を信号として与えることによって学習を行う．それにより，類似盤面に対して汎化を行うことができる．この手法を実装することにより，異なる対戦相手に

対し勝率をあげることに成功したことを実験結果により示す.