3M-4

強化学習における状態統合による学習の高速化

長谷川 雄吾 † Yugo Hasegawa
宮内 新 † Arata Miyauchi
荒井 秀一 † Syuichi Arai

1. はじめに

強化学習とは, 学習エージェントが環境とのインタラクションによって得た報酬をたよりに方策（観測状態から行動への写像）を獲得する機械学習の一つである. 事前に学習する環境の知識を必要としないため, 未知の環境やゆるやかに変化する環境に対応することができる.

強化学習では一般的に状態空間を離散化し, 離散化された各状態に対して方策を学習していく. 細やかな行動制御のため離散化を細かく, もしくは状態入力の次元数が増加すると状態数は指数関数的に増加し, 学習に必要な計算量やメモリ量が非現実的なものとなってしまう.

実際の環境では, 状態入力は多次元かつ細かな離散化が必要になることが多い. 実用的なタスクを学習させるためには, 計算量が状態数の数に大きく影響を受ける問題を解決する必要がある.

2. 研究の目的

計算量を減らすための単純な方法は, 状態が細かくならないように状態空間の分割を設計し状態数を少なく抑えることである. 状態の分割を粗くすることで不完全知覚状態が増加しやすく, 獲得された方策によって得られる行動系列も不安定なものになりやすい.

逆に不完全知覚状態数を最低限に抑えるためには, 有効な行動が異なる部分空間を分割するのに十分な細かさの状態が必要となる. このとき細かな状態の分割が必要ない部分空間において, 学習の効率が低下し計算量が増加してしまう.

以上のように, 細やかな行動の学習と学習に必要な計算量の削減のあいだには, トレードオフが存在する. 本研究では, 細かな状態分割を行ったうえで学習の効率を下げている部分の状態をまとめ, 集合として扱うことで, このトレードオフを解消することを目指す.

3. 提案手法

本来同一の状態としてみなすべき状態群を, 一つの状態集合として扱い, 学習の効率を向上させる.

本研究では, 環境のもつ構造によって同一の行動が同様の結果を引き起こす観測状態群を, 本来おなじ状態としてみなすべき状態として集合を構成し行動選択に利用する. 行動選択は確率的なもの（非グリーディー) であれば, その方法を問わないが, 今回はルーレット選択 (獲得した報酬の大きさに応じた選択確率で選択) を用いた. また, 学習には ProfitSharing を用いている.

各観測状態毎に学習した場合, 獲得した報酬が分散されるため学習効率が低下する. このとき行動選択に集合を用いれば, 分散した報酬をまとめることができる. 各観測状態毎に学習した場合に比べ, 環境とのインタラクションを行った回数が同数であっても, 状態当たりの多くの学習回数を経験した場合と同様の効果が期待できる.

†武蔵工業大学 Musashi Institute of Technology

4. 状態集合の生成

複数の観測状態が本来同じ状態とみなせるかどうかを, 次の二つの条件を満たすかどうかで推定する.

- 集合と観測状態において同一の行動が最大の報酬値を持つ (条件 1)
- 集合に属する観測状態から一行動で遷移可能であるか (条件 2)

条件 1 を満たす場合, 対象下の状態で最も効果の高い行動が同じである可能性が高い.

どの行動が最も効果の高い行動になるかは, 対象とする状態での環境の持つ構造が決定する. 迷路問題の場合であれば, どの方向にゴールが存在するか, 壁はどこに存在するかといったことが環境の構造になる.

条件 2 によって, 条件 1 を満たす状態群の中から, 最大の報酬を持つ行動が同一であっても, 環境の構造が異なる観測状態を除外する.

これは, 環境の構造が異なる場合, 学習が進行すると異なる解を持つことがあり, 得られた解の質が低下することを防ぐためである.

学習エージェントが報酬を得た場合, 報酬を得た観測状態から遷移情報をたどり条件 1, 条件 2 を満たす状態群を探索する.

次に得られた状態群の報酬の総和を各行動毎にとり, 状態集合での報酬の分布とする. 次のインタラクションから学習エージェントは, 集合に含まれる観測状態を観測したとき, 状態集合の報酬の分布を使い行動を選択する.

5. 実験設定および結果

提案する手法を検証するため,5.1 迷路問題,5.2 追跡問題での学習を行った.

すべての実験において行動選択器にはルーレット選択を, 学習器には ProfitSharing (割引率=行動数) を用いて実験を行った.

5.1 迷路問題

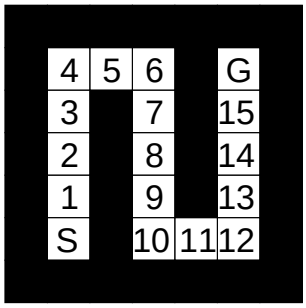


図 1: 迷路問題

提案する認識器を用いた場合, 用いない場合の 2 パターンで学習を行った. 学習曲線を図 2 に示す. 縦軸がゴール到達までの行動数 (最短 16), 横軸が学習回数を表す.

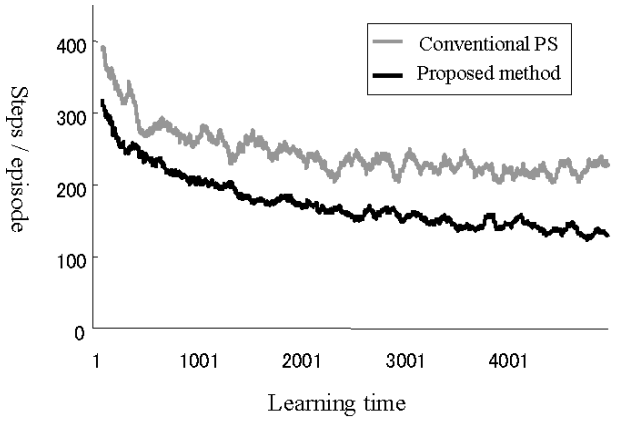


図 2: 迷路問題学習曲線

提案する認識器と組み合わせることによって学習の効率が向上し学習時間の短縮が達成できた. 状態空間の構成は,PS との組合せでは S・1・2・3・4-5・6-9・10-11・12-15 がそれぞれ一つの状態となっていた.

5.2 追跡問題

強化学習のベンチマークとして一般的な追跡問題で実験を行った. 環境の広さは 5x5 のトーラス空間とし, この中で二体の学習を行うハンターエージェントがターゲットを捕獲するための行動を学習する. 選択できる行動はハンターが上下左右の四方向, ターゲットはこれに加え斜め四方向の八方向とした.

迷路問題と同様に提案する認識器を用いた場合用いない場合で学習を行った. 学習曲線を図 3 に示す. 縦軸がゴール到達までの行動数, 横軸が学習回数を表す.

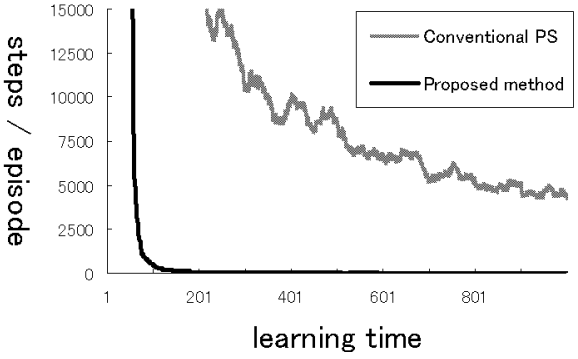


図 3: 追跡問題学習曲線

状態空間は提案する認識器と組み合わせた場合, 行動の種類毎 (上・下・左・右・停止) に対応し状態数が 5 つになるまで統合が進んだ. さらに学習回数を重ねていくとおよそ 3000 エピソード程度で 5 つの行動のうち最も効果が低いと考えられる停止行動に対応した状態が, 他の行動に対応する状態に吸収され 4 状態に収束した. (図 4)

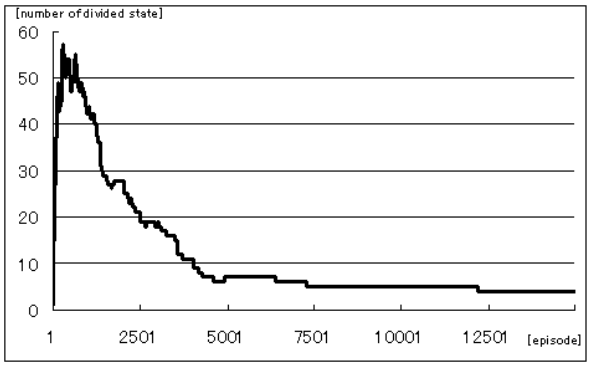


図 4: 追跡問題状態分類結果

提案する手法は, 状態数が行動の数にまで統合されることが期待できるが, この結果はあまり有効でない行動が行動集合の中に含まれていた場合それを除いた行動数にまで収束できる可能性を示しており, 好ましいと言える.

6. 考察

迷路問題, 追跡問題において状態数の削減・学習時間の短縮・状態空間の自律的設計を行うことができ, 環境の構造の推定に基づき状態を自律的に構成することで学習を行うのに有利な状態空間の設計が行えることを示せた.

とくに状態遷移の複雑な問題 (追跡問題) がそうでない問題に対して効果が大きいことがわかる. これは複雑な問題環境を学習可能にするために有効な性質であると考えられる.

7. 今後の課題

現状では, 異なる状態を同一の状態として認識してしまう不完全知覚が存在する環境では, 一方の状態がある状態集合に属することで, 他方の行動選択において誤った行動を選択し続ける恐れがある.

これに対応するために, 誤った状態集合に統合されたことを検出し, 行動選択に状態集合を利用しないようにする仕組が必要である.

参考文献

[1] 木村 元 ,L. P. Kaelbling, 部分マルコフ決定仮定下での強化学習 J. JSAI, vol. 12, no. 6, pp. 822-830, 1997

[2] 宮崎 和光, 荒井 幸代, 小林 重信, POMDPs 環境化での決定的政策の学習, J. JSAI, vol. 14, no. 1, pp. 148-156, 1999 (In Japanese)

[3] 宮崎 和光, 山村 雅幸, 小林 重信, 強化学習における報酬割り当ての理論的考察 J. JSAI, vol. 9, no. 4, pp. 580-587, 1994