アフォーダンス理論に基づく状態空間の自律的構成

長谷川 雄吾 †
武蔵工業大学大学院 §

宮内 新 ‡
武蔵工業大学大学院 §

1. はじめに

強化学習とは問題環境と学習エージェント (行動主体) の相互作用 (認識, 行動, 評価) により, 学習エージェントが目的とするタスクを達成しうる方策を獲得する機械学習の枠組の一つである.

一般的に学習エージェントは次の 3 つの機能モジュール, 状態の認識を行う認識器, 行動を選択する行動選択器, 学習を行う学習器から構成される. 本研究では強化学習における認識器の状態空間の自律的構成を主題とする.

2. 研究の目的

強化学習を適用する問題が mdps に代表されるような比較的単純な問題環境から, より複雑で実際の問題環境に近いものへと拡張されてきた. この結果次のような問題が大きな影響を持つようになってきている.

- 状態数の爆発的増加
- 学習時間の増加
- 状態空間設計の困難さの増大

これらの問題に対応するために状態数の増加抑制および学習効率の向上が必要になる. そこで本研究では状態空間を自律的に構成し状態数を最小限に抑え, 状態の汎化による学習効率の向上が可能となる状態認識を提案する.

3. アフォーダンス理論に基づく認識

強化学習は環境との相互作用によってタスク達成に貢献した行動を強化することで方策を学習する. この点で強化学習はアフォーダンスを抽出する過程であると言える. そこで新しい認識の機構として, アフォーダンス理論に基づく認識を行う認識器を提案する.

アフォーダンスは環境の構造によって決定する. 環境の構造によって行動の価値が定まりその行動がどんな効果を持たすか (アフォーダンス) が与えられている. 提案する手法では, 感覚入力 (センサ) の値, 行動による状態の遷移, エージェントが得た報酬を用いて環境の構造を推定し状態空間を自律的に構成していく.

Autonomous Construction of State Space Based on Affordance Theory
†Yugo Hasegawa
‡Miyauchi Arata
§musashi institute of technology

報酬に基づいて状態を分類することで複数存在するアフォーダンスのうちでタスク達成に貢献する行動毎に分類することができ学習に適した分割が可能になる.

提案する状態認識器では, つぎの機能が必要となる.

1. センサの値からの状態認識
2. 行動後の状態遷移の記録
3. 報酬と遷移記録からの状態空間分割

以下順に説明する.

3.1 センサの値からの状態認識

センサの値は学習エージェントが環境の状態を推定するための, 唯一の手掛かりである. 現状態までの遷移を推定に利用することも考えられるが, ここでは短にセンサの値が幾何的にもっとも近いものを学習済みの知識から探し, その学習結果が属する状態を仮の現状態として認識する.

3.2 行動後の状態遷移の記録および状態の作成

仮の現状態の学習結果を用いて行動を選択実行する. 実行の結果, 状態遷移を観測し仮の現状態と矛盾があれば, センサの値と遷移先の情報を新しい状態として定義する.

3.3 報酬と遷移記録からの状態空間分割

報酬を得た状態から状態の遷移によってリンクした学習結果を調べていき, 同一行動が最大の報酬値を持つとき, タスク達成に対して同一の環境の構造を持つと考え同一の状態として統合する.

4. 実験設定および結果

提案する手法を検証するため,4.1 迷路問題,4.2 追跡問題での学習を行った. 現時点では 3.1,3.3 を実装し,3.2 については知識に一致するセンサ情報がない場合は, 必ず毎回新しい状態として定義して実験を行った.

両実験においてともに行動選択器にはルーレット選択を, 学習器には ProfitSharing および PS を改良した DPS[1][2] を用いて実験を行った.

4.1 迷路問題

図 1 のような簡単な迷路で実験を行った. エージェントが選択できる行動は上下左右の 4 行動で, 状態の認識は全てのマスを区別して行う. 壁の有無は認識できず壁方向に行動した場合, 状態の遷移は生じない. PS の割引率は 4 とした.

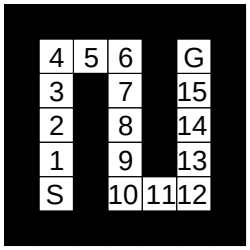

図 1: 迷路問題

提案する認識器を用いた場合用いない場合をそれぞれ PS・DPS と組合せ学習を行った. 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.

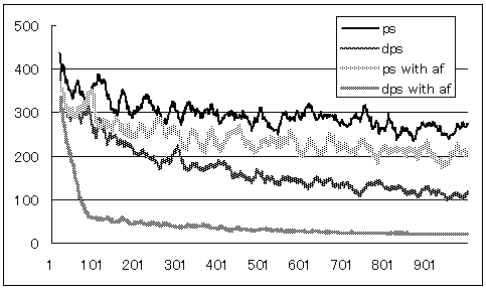

図 2: 迷路問題学習曲線

PS・DPS ともに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. DPS との組合せではゴールから離れたマスにも学習の結果報酬が行きわたるため S-3・4-5・6-9・10-11・12-15 のように状態が構成されスタート付近での統合が進んでいる.

4.2 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.

エージェントが選択できる行動は上下左右と停止の 5 行動で, 状態の認識は視界内 (5x5) の獲物と見方ハンターとの相対位置によって行う. 獲物は全ハンターからの距離の総和が最も大きくなるように行動する. PS の割引率は 5 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をそれぞれ PS・DPS と組合せ学習を行った. 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.

状態空間は PS・DPS ともに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

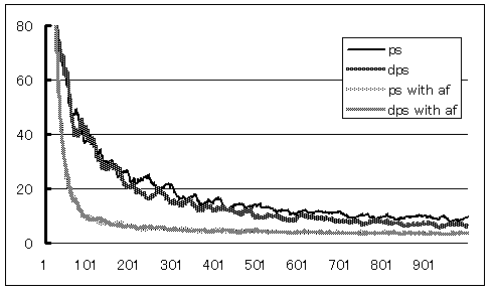

図 3: 追跡問題学習曲線

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. 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.

5. 考察

MDPs 環境, 非 MDPs 環境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.

6. 今後の課題

実験から提案する認識手法の基本となる考え方を用いることで, ある程度の性能向上が期待できることが示せた. そこで未実装である**??**の仮の状態の確認を実装し, 汎化能力を持たせることでさらに学習効率をあげ, さらに複雑な環境での学習を可能にしていくこと. また実装後どのレベルまで学習可能となるかを検討すること等を今後の課題とする.

参考文献

[1] 長谷川 雄吾, 高田 沙都子, 宮内 新, 荒井 秀一 : Profit Sharing を改良したより効率的な強化学習手法 (1)-選択確率による報酬割引率決定手法- , 情報科学技術フォーラム Vol.FIT 2003, 情報技術レターズ Vol.2, Page125-126

[2] 高田 沙都子, 長谷川 雄吾, 宮内 新, 荒井 秀一 : Profit Sharing を改良したより効率的な強化学習手法-Dynamic Profit Sharing での合理性の検討- , 情報科学技術フォーラム Vol.FIT 2003, 情報技術レターズ Vol.2, Page127-128