ISS-P-127

鏡像を利用したQ学習の高速化

北尾　健大　　三浦　孝夫
法政大学理工学部創生科学科

1. まえがき

環境と連動しながら、対処方法に関する知識を自動的に得ることができれば、多様な問題への効果的な解を効率よく得るができる。ここでは、動作主(エージェント)と環境だけではなく、複数の動作主間の協調も含まれる。

目標に対して動作主がこれを追跡し捕獲するために取るべき戦略を選択する問題を、追跡問題という。多くの場合、決定的な解が無いため、経験から戦略を選択する方策を抽出したい。この学習をおこなうため、本稿では強化学習である Q 学習 を用いる[1]。Q 学習は、予め正解例を与えないため多様な状況に柔軟に対応できるが、学習速度が遅く、また抽出される戦略の信頼性(精度)の保証がない。本研究では、Q 学習を高速化するため、鏡像手法を提案し、学習速度および精度の評価を行う。

2. Q 学習

Q 学習では実行するルールに対し有効性を示す Q 値という値を持たせ、動作主の行動に従って値を更新する。ルールとは状態とその状態下の行動を対にしたものである。Q 学習は以下の更新式で表現される。

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a' \in A(s')} Q(s',a') - Q(s,a)]$$

s は現在の状態、a は s のときにとった行動、s'は次の状態、a'は次の状態の行動である。Q 学習は全状態を十分に繰り返し学習したとき最適な評価値に収束する。

3. 追跡問題

グリッド空間でハンタが獲物を捕獲する知識を検出することを追跡問題という。ハンタと獲物は各一体以上存在する。ハンタは視覚を持ち、自分を中心に獲物を相対的に知覚する。ハンタは行動選択手法で上下左右停滞の五つの行動を選択。獲物は同じく5つの行動をランダムに実行する。捕獲条件はハンター体の場合は獲物と隣あったとき。二人のときは二方向を塞いだときとする。

4. 提案手法

本研究ではグリッド空間を二次元座標と考え、空間の中心に X 軸、Y 軸、原点において対称のハンタと獲物を仮想的に設ける。それら鏡像のハンタ、鏡像の獲物と呼ぶ。実物が移動するとき鏡像も対称性を伴って行動する。このとき実物と鏡像の Q 値の同時更新を提案する。

Q 値を同時更新する際、(状態、行動)の組み合わせが同じ Q 値を更新する際干渉が生じ精度が悪化する。

5. 実験

本研究ではX軸、Y軸、原点対称の3つとそれら組み合わせの合計7個のパターンを調べる。ベースラインとして通常の Q 学習を用いる。評価基準は学習中の収束したときの最終値、Q 値の収束時間、学習後の Q 値を利用した捕獲までの回数とする。

グリッド環境を14×14に設定する。ハンタが一体のとき視野は 3×3、7×7、11×11。ハンタが二体にとき 7×7,11×11,15×15 で実行。ハンタと獲物が配置されてから捕獲までを1エピソードと呼ぶ。ハンター体のとき学習中、学習後共に200000回で実行。二体とき学習中400000回(15×15のみ600000回)、学習後200000回繰り返す。

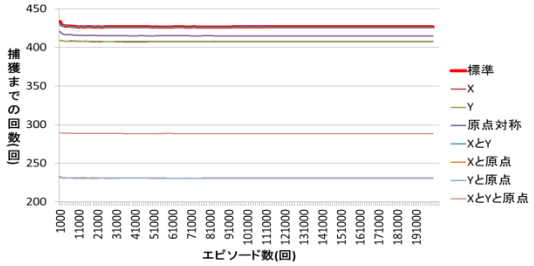

図 1 ハンター体、視野:3×3 の学習経過

ハンター体の場合学習中の収束したときの最終値はベースラインと比較して視野が狭い順に最大 45.9%、39.3%、30.0%減になる。学習後の捕獲回数には変化が見られない。収束時間は視野 3×3 のとき悪化して最大 6.6 倍遅くなる。5×5 と 7×7 は向上して最大 7.2 倍、3.5 倍になる。

ハンタが 2 人の場合，同様にベースラインと比較して収束したときの最終値は視野が狭い順に最大 39.1%、24.8%、21.7%減になる。学習後の捕獲回数も、11.1%、9.0%、5.3%減少する。収束時間は 7×7 は 3.1 倍。他は差がない。

注目すべき点は収束速度の違いである。鏡像の数を増やし視野が狭いことでの鏡像干渉の増加が原因である。

6. 結論

収束したときの最終値は視野が狭まるほど減少の割合増加した。ハンタが二人の場合、学習後の捕獲回数も減少した。鏡像干渉は視野が狭いほど生ずる。本研究で提案した鏡像手法は視野の広いものに対しては有効だと考えられる。今後の課題として視野の大きさに応じた精密な測定による鏡像干渉の影響を調査が考えられる。

参考文献

[1] 高玉圭樹“マルチエージェント学習―相互作用の謎に迫る―,” コロナ社(2003)