卒業論文要旨

知覚情報の利用選択によるマルチエージェント強化学習の効率化

(知能情報システム学)　　山本　健弘

１．緒言

マルチエージェントシステムに強化学習を用いる場合、エージェント数の増加に従い指数関数的に状態数が増え、それに伴う学習時間の増加がボルトネックとなっている。そこで「ハンター問題」を例に、与えられる情報の全てを利用するのではなく、重要な情報を選択し利用することにより状態数を削減し、学習を効率化する手法の提案を行う。

２．開発環境

OS は Microsoft 社 Windows2000 であり、プログラミング言語は C++である。また、開発環境として Microsoft 社 VisualC++6.0 を使用した。マシンは VAIO PCV-RX71K、CPU は IntelPentium4 1.5GHz、メモリは 512MB を使用した。

３．方法

３-１　ハンター問題とは

ハンター問題とは、1986 年に Benda により初めて導入され以後様々な研究者が問題設定を変え使用している問題で、協調型マルチエージェント強化学習の性能評価をする上での標準問題である。

具体的には、図１(ｎ＝７、ハンター数３の場合)のようにｎ×ｎマスのトーラス状グリッド空間で、動き回る獲物（●）を複数のハンター（■）が捕獲するというタスクであり、獲物とハンターは１ステップ毎に上下左右のうちの１方向に１マスだけ移動するか、停止することができる。また「捕獲」については、ハンターが獲物と同じマスに移動した時とする場合と、複数のハンターが獲物を取り囲んだ時とする場合があるが、本論文では後者を採用した。

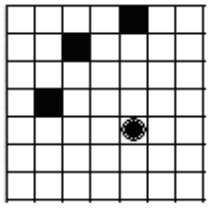

図１　ハンター問題

表１　ハンター問題の状態数

	3	5	7	9	11	13	15
1	9	25	49	81	121	169	225
2	81	625	2401	6561	14641	28561	50625
3	729	15625	117649	531441	1771561	4826809	1.1E+07
4	6561	390625	5764801	4.3E+07	2.1E+08	8.2E+08	2.6E+09
5	59049	9765625	2.8E+08	3.5E+09	2.6E+10	1.4E+11	5.8E+11

３-２　強化学習について

強化学習とは、数値化された報酬を最大にするために、どの行動を選択すればよいかを学習す

るものである。状態価値関数とは、その状態の望ましさを表す関数であり、状態の望ましさとは、その状態から将来得るであろう期待報酬である。

状態価値関数にさらに行動を含めた表現が行動価値関数である。行動価値関数 Q(s,a)は、方策πにおいて行動 a をとり、その後は方策πに従って行動したときの期待報酬である。Q 値の更新は以下の式で表される。

$$Q(S_t,a_t) \leftarrow Q(S_t,a_t) + \alpha(r_{t+1} + \gamma \max_a Q(S_{t+1},a) - Q(S_t,a_t))$$

ここで、$\max_a Q(S_{t+1},a)$は状態 S_{t+1}で最も行動価値関数の高い行動 a における行動価値関数で、最適行動価値関数と言う。また、αは状態価値関数の更新の度合いを示すパラメタで学習率と呼ばれ、$0<\alpha \leqq 1$で、γは割引率と呼ばれ将来の報酬が現在どれだけ価値があるのかを表すパラメタで、$0 \leqq \gamma \leqq 1$である。このようにQ値を用いた強化学習をQ(０)学習と言う。

本論文では強化学習アルゴリズムとしてQ(０)学習を用いた。

３-３　状態数削減の必要性について

強化学習により最適政策へ収束するには、全ての状態での全ての行動について十分な経験回数が必要となる。つまり、状態数が爆発的に多い状況では計算時間も膨大になり実用的ではない。

表１は、縦をハンター数、横を格子数として状態数を表している。例えば、グリッド数が７×７、ハンター数が４の場合、あるハンターから見た状態数は7^8＝５,７６４,８０１である。さらに行動が５種類あるので１ハンターにつき２８,８２４,００５種類のQ値が存在することになる。この値は現在の計算機にとってもかなり大きいものである[1]。

３-４　情報の利用選択について

ハンターが知覚できる情報には、様々なものがある。しかしながら全ての情報を使用したのでは、上記のような状態数の爆発問題が起こってしまう。そこで、学習する上で特に重要な状態であるかそうでないかを学習する際に、あるエピソード毎の最大Q値により判断させ、さほど重要でない状態では特に必要であろう情報のみを選択し使用することで状態数の削減を図った。

具体的にハンター問題では、学習する上で重要である状態では「方向」「距離」共に利用するが、学習する上でさほど重要ではない状態では、「方向」のみを利用する。また、「方向」についても正確な「方向」ではなく上下左右の４方向に祖視化して利用する。重要かどうかの判断については、数エピソード毎で獲物との相対位置のみでの最大Q値の最大値(maxQmax)をとり、maxQmax が基準となるQ値１(stdQhigh)より大きければ隣のグリッドも重要であると判断し、maxQmax が基準となるQ値２(stdQlow)よりも小さければそのグリッドをさほど重要でないとする。結果については卒業研究発表会で報告する。

参考文献

[1]伊藤　昭、金渕　満、” 知覚情報の粗視化によるマルチエージェント強化学習の高速化 ”電子情報通信学会論文誌、Vol.J84-D-I、No.3、pp285-293、2001.