5Q-3

Q学習による知覚情報の粗視化による追跡動作の学習

桑原　直哉†　　三浦　孝夫‡

法政大学工学部情報電気電子工学科

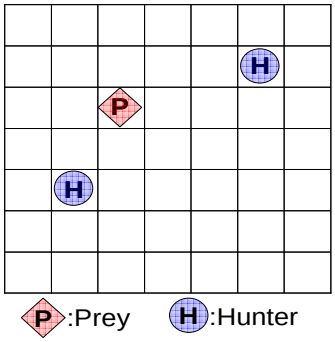
東京都小金井市梶野町３－７－２

１．　前書き

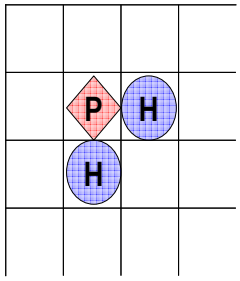
近年、自律的エージェントの行動を解析することで、複雑で大規模な問題を解決する研究が進められている。しかし複数の自律的エージェントが協調して（予め設定された）目標を達成するのは容易ではない。そこで解法の一つとして強化学習が提案されるが、状態数の増加に伴い学習時間の遅さ、記憶領域の膨大が問題となっている。本論文では、マルチエージェント系におけるベンチワークの一つである追跡問題を通して、行動しながら学習を同時に行うQ学習によって、知覚情報の粗視化手法を提案し、追跡動作の効率化を行う。ここでいう知覚情報の粗視化の手法とは、近くの位置は詳細に把握できる知覚情報のひな形を使用しエージェントのいる方向に平行移動させ、遠くの位置に対しては方向のみで知覚する手法である。

２．　追跡問題

本論文では以下の設定に基づく追跡問題を扱う。



〈図１．追跡問題〉



〈図２．目標状態〉

ｍ×ｍ格子状の環境を設定し、２つの追跡者(Hunter)エージェントと１つの獲物(Prey)エージェントを配置する。初期状態から目標状態に達するまでを１エピソードとすると、エピソードごとに初期配置は無作為に与える。各エージェントは同時に行動し、上下左右の方向に１マス進む。もしくは停止の５つの行動の中から一つを選択する。また獲物エージェントは無作為に行動を選択する。各エージェントが同時に行動を終えるとき１ステップとする。各エージェントがある行動をしたときに同じマスでエージェント同士が衝突した場合、そのエージェントは１ステップ前の位置に戻る。また格子の壁は越えることができないと仮定する。

各エージェントの視界は全範囲のマスを見ることができる。本論文では追跡者エージェント同士または獲物エージェントを知覚する場合は相対位置で情報を保存する。

目標状態は図２のように、２人の追跡者エージェントが獲物エージェントに対して隣接する場合とする。ここで「隣接する」とは、上下左右のいずれかに位置する場合に隣接する状態とする。

３．　状態数削減の必要性

強化学習は観測可能なすべての状態Ｓと可能な行為 a との対(Ｓ, a)について行動の評価を行い、後に同じ行動に至ったときはその行動評価値を用いて行動選択に利用する。強化学習が効果を発揮するためには、可能な状態・行為の組(Ｓ, a)のすべてについて統計的な情報を収集できる十分な回数を経験することが必要になる。そのため状態数が膨大な環境では十分な回数を経験することができず、状態数が多いために時間がかかり学習速度を落とし、実用的ではない。

	m = 3	m = 5	m = 7	m = 9
n = 1	9	25	49	81
n = 2	81	625	2401	6561
n = 3	729	15,625	117,649	531,441
n = 4	6,561	390,625	5,764,801	43,046,721

〈表１．格子数m、エージェント数ｎの状態数〉

ここで追跡者の視界をｍ、エージェント数をｎとする。この表のように複数のエージェント環境下では、エージェント数の増加に伴って指数関数的に状態数が増える。

４．　Q学習

本研究では強化学習の中からQ学習を使用する。Q学習は、期待報酬の推定値である規則の価値を他の推定値に依存して更新するものである。Q学習の更新式は(1)である。

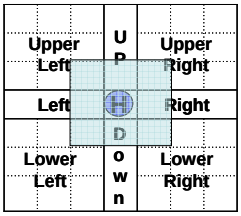
$$Q(s,a) \longleftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a' \in A(s')} Q(s',a') - Q(s,a) \right] \quad (1)$$

Q(s,a)は状態 s において行動 a をとるときの価値、Q(s’,a’)は遷移先の状態 s’において行動 a’をとるときの価値を示している。また本論文で使用する行動選択手法は、確率的政策である Boltzmann 分布を用いて選択する。

$$p(a \mid s) = \frac{e^{\frac{Q(s,a)}{T}}}{\sum_{a_i \in A} e^{\frac{Q(s,a_i)}{T}}} \quad (2)$$

５．　Q学習と粗視化

本研究では、Q学習を用いた知覚の粗視化を提案している。複数のエージェント環境下のときに生じる状態数の爆発を防ぐ必要がある。本稿で提案するアイディアは、



〈図３．粗視化手法〉

図３のように近くの位置には詳細に位置を把握できる知覚情報のひな形と、遠くの位置には方向でしか分からない知覚器を作り、獲物がいる方向に対してそのひな形を平行移動することで、状態数の削減を行う手法である。これによって遠くの位置であるときは目標達成にさほど影響がないため粗く方向だけ知覚し、一方自エージェントの近くは目標達成のための体系を把握し易いために詳細に知覚をする。また、この提案手法により状態数の削減に２エージェントの場合１６×１６×５(５方向)の状態数だけに削減ができる。

文献[2]では Expansion の手法を使っている。これは、段々と粗視化から詳細に知覚できるようになっているが、粗視化したときのQ値を詳細に知覚するときに利用するとき、知覚の粗視化をしていたときのQ値によって、詳細に知覚したときに悪影響が起こることがある。最終的にすべてを知覚できるようにしているためにさほど状態数の削減に繋がってはいない。本研究は文献[2]を発展させた手法を提案する。本手法により更に広い環境でも知覚の粗視化により、状態数は変わらないという利点がある。このことよりQ学習と粗視化を組み合わせてQ学習の粗視化を提案する。

６．　実験と考察

この実験では 10 万エピソードの繰り返しを１０回試行するときの学習を検証する。表１では１万エピソードから 10 万エピソードまで１万ずつ区切り、その後の 100 エピソードの平均捕獲ステップ数を表している。また学習初期の状態として 2500, 7500 回目の状態を用いる。

パラメータ初期値は学習率 α=0.04、割引率 γ=0.9、T=0.2、目標達成のときの報酬 r=10、Q値の初期値は 1.0　とする。学習率 α は実験によりこの数字が一番いい結果を出るためこれを用いる。

各学習後の１００エピソード		
エピソード	１０施行の平均ステップ数	平均分散
2500回目	7.4749	10.0899
７５００回目	7.2083	2.7099
１００００回目	7.2679	2.2201
２００００回目	7.0233	3.6237
３００００回目	6.3219	0.7225
４００００回目	6.6145	1.9770
５００００回目	7.0669	0.7615
６００００回目	6.7591	1.3453
７００００回目	6.6214	1.1067
８００００回目	6.1747	2.4554
９００００回目	6.6993	1.2815
１００００００回目	6.9642	1.8996

〈表１．各学習後の平均ステップ数〉

表１より、１０万エピソードを１０試行繰り返したときの平均捕獲ステップ数は学習初期と学習後期と比べた場合、少しだが学習後期のほうが少ない。これはQ学習によって学習した成果, つまり初期段階から学習してステップ数が少なくなっているといえる。これは本研究のQ学習の粗視化手法を用いたことで状態数が減り、いち早く統計的な回数を重ねられたことがステップ数に表れていることが分かる。平均分散が初期状態は大きくなっているが学習後期になると差がステップ数以上になっている。これも統計的な回数を重ねることで捕獲するにあたって適切な行動を移すように淘汰されている。

７．　結論

本研究ではQ学習の知覚情報の粗視化を行い、状態数の削減をした。今後は、より学習性能が劣化しないよう、より複数なエージェントの場合やより広い環境下で粗視化手法を検証する必要がある。

[参考文献]

[1]荒井 幸代、宮崎 和光、小林 重信：マルチエージェント強化学習の方法論 - Q-learning と Profit Sharing による接近 -、人工知能学会誌 Vol.13,No.5,pp.609 - 618,1998

[2]金重 徹、片山 謙吾、南原 英生、成久 洋介：知覚情報の粗視化に基づくマルチエージェント強化学習の性能比較、自律分散システムシンポジウム、2007