追跡問題における報酬割り当て方法に関する一考察

○辻和輝　植村渉（龍谷大学）

About Credit Assignment for Multi-agent Security System

* K.Tsuji and W. Uemura (Ryukoku University)

Abstract－ In recent years, the security industry focuses on the patrol security robot system. In the case of security system with multi-robots, it is necessary for multi robots to consider the other robot action in order to reinforce their action selections. The Episode-based Profit Sharing (EPS) can assign the reward for multi-agents from the view of the rationality. The reinforcement function of EPS cumulates the value to its rule by the reward. In this paper, we focus on this reinforcement function and propose to adapt the update function to the reinforcement function of EPS.

Key Words: Reinforcement Learning, Profit Sharing, Multi Agent

1　はじめに

近年，ロボット技術の発展に伴って社会へのロボットの導入が進んでいる. 施設等の警備業務においては，閉鎖されている休日や夜間といった期間に警備を行う. 警備業務では人力による業務の割合が高いため，労働者である警備員の負担が大きくなる問題がある. また，監視カメラを用いた手法では，視野角の制限から死角が発生し, 施設内の全域を監視することが困難となる.

そのため，警備業務では移動式ロボットを導入して施設内の巡回を行うことで, 業務の改善を図っている. 巡回警備を行う移動式ロボット（巡回警備ロボット）には撮影用のカメラを取り付けて，施設内の一定区間を自律的に巡回する．巡回警備ロボットは不審者の発見等，異常の検知をした場合に遠隔の監視オペレータへカメラ画像を送信することで報告を行うと共に不審者を追跡する．しかし人間とロボットでは状況認識能力に差があるため，単独のロボットで不審者を追い詰めることは困難であり，複数の移動式ロボットが協調する必要がある.

複数のロボットが同時に存在している場合，各々のロボットが自身にとって最適な行動のみを選択し続けると，ロボット全体において合理的ではない場合がある．そこで本研究では，強化学習の一つである，Profit Sharing を侵入者追跡に用い, その強化方法について議論する.

2　強化学習

強化学習の枠組みではエージェントと呼ばれる学習者と環境と呼ばれる学習を行う空間が相互作用を行うことによって，エージェントが状況に応じて適した行動を獲得していく．エージェントと環境の間には，エージェントの置かれている状況を表す状態 s ，エージェントが状態に応じて選択，実行する行動 a ，環境から与えられる行動の良し悪しを表す報酬 r ,状態と行動の対であるルール(s, a)の価値を表現する Q 値がある. エージェントは時刻 t において状態 s_t を観測する. 知覚能力が完全でない場合には, 観測として o_t を用いる. そして状態 s_t で選択可能な行動群から行動 a_t を行動選択方法に基づいて選択し，それを実行する．エージェントが行動を実行することで環境内でのエージェントの状態が次状態 s_{t+1} へと遷移し報酬 r_{t+1} を受け取る. 報酬は遷移した次状態がエージェントの目標状態であれば正の報酬値を与え，それ以外では 0，または負の報酬値を与える場合が多い.

受け取った報酬からエージェントはルールの価値である $Q(s_t,a_t)$を更新する．また，エージェントが環境全体を観測できる場合，エージェントがマルコフ性を満たしていることからマルコフ決定過程（Markov Decision Process，以下 MDP）と呼ぶ．MDP 環境下ではエージェントは環境を完全に知覚できるため，観測情報を示すパラメータ o_t は $o_t = s_t$ である．実機のロボットが動作する現実的な環境では，センサのノイズや誤差等の影響や測定限界から環境すべてを知覚できない場合が多い．このようにエージェントの知覚能力が制限されており，問題空間を部分的にしか観測できない場合は部分観測可能マルコフ決定過程（Partially Observable Markov Decision Process，以下 POMDP）と呼ばれる．POMDP 環境下では本来異なる複数の状態を，エージェントが同一の状態として観測する場合に問題が発生する. POMDP では状態の代わりに観測 $o(s_t, a_t)$を用いて学習を行う [2]．観測情報 o は状態 s の部分的な観測値である．ここで，報酬獲得に必要な行動が異なる状態群を同一観測として得る場合，その観測において選択すべき行動が複数生じる．このように本来別の状態を異なって同一と観測する際の問題を不完全知覚問題と呼ぶ. MDP 環境に対する強化学習法では, 同一観測（状態）に対して，一つの行動のみの強化を前提としているため，不完全知覚問題が生じる POMDP 環境では，適切に学習が進まない.

2.1 学習手法

強化学習の一手法である Q-learning[1]では時刻 t において選択したルール(s_t, a_t)の価値 $Q(s_t, a_t)$を式（1）で更新する．

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a_t) - Q(s_t, a_t) \right] \quad \cdots (1)$$

式（1）の Q 値の更新は，推定値ベースの更新である．あるルールを実行した際に獲得する報酬を価値の推定値とし，そこから将来に獲得する報酬を割引期待報酬値として扱い，Q 値がその値になるように更新を行う．この式においてγは割引率を示しており将来に獲得する報酬をどれだけ重視するかを表している．$0 \leq \gamma \leq 1$の範囲を取り，0 に近いほど即時報酬を重視し，1 に近いほど将来の報酬を重視する．また，α は学習率であり，$0 \leq \alpha \leq 1$の範囲で設定する．式（1）の右辺第一項は更新前のルールの Q 値，第二項は割引期待報酬値である．ここでα を 0 に近づけると更新前の Q 値を重視し，1 に近づけると割引期待報酬値を重視して Q 値を更新することになる．

一方 Profit Sharing[3]での Q 値の更新方法は，報酬を獲得したときに，エピソードと呼ぶ目標状態へ遷移するまでの行動系列を遡って報酬を分配する．Profit Sharing の更新式は式（2）である．

$$\omega(s_x, a_x) \leftarrow \omega(s_x, a_x) + r \times f(x)$$
$$Q(s_t, a_t) = \omega(s_t, a_t) \quad \cdots (2)$$

あるエピソード内に同一のルールを複数回選択している場合，その間の行動系列はループの形をとる．このループ上の行動系列を迂回系列 [4]と呼び，迂回系列上のルールを無効ルールと呼ぶ．無効ルールは報酬獲得に対して貢献しない．それ以外のルールを有効ルール呼び，報酬獲得に貢献したルールである．更新式(2)における強化関数 $f(x)$ は報酬値の割り当てに用いるため報酬分配関数と呼ぶ．ここで x は報酬獲得時からさか遡るため，エピソードの時系列と逆向きの変数となる．MDP 下では強化関数として等比減少関数を用いることで無効ルールの強化を抑制することが知られている．等比減少関数では選択可能な有効ルール数 L を用いて等比を$1/L$で表す．

$$f(x) = {}^1\!/_{L^x} \quad \cdots (3)$$

PODMP 下では不完全知覚の影響により，現在の観測と行動選択後の観測の関係が一致しない場合が生じる．Q-learning の更新式では更新において割引期待報酬に時刻 t+1 の次状態を用いて計算するため，不完全知覚の影響を受ける．Profit Sharing では更新式に時刻 t +1 の次状態を含まないため，不完全知覚の影響を受けないが，報酬分配関数 $f(x)$で用いる等比減少関数にて時系列の値が含まれるため，不完全知覚の影響を受ける．そこで報酬分配関数から時系列の項を取り除いた関数を用いる Episode-based Profit Sharing （EPS）[5]がある．EPS ではエピソード内で選択したルールに対して均等に報酬を分配することで不完全知覚下においても影響を受けず，学習が進むことが知られている．式（4）が EPS で用いる報酬分配関数であり，W はその分配時のエピソード長である．

$$f(x) = {}^1\!/_{L^W} \quad \cdots (4)$$

2.2 行動選択方法

エージェントは，行動選択の判断にルールの価値である Q 値を用いる．Q-learning ではルールの価値の推定値に近づくように Q 値を更新するため，報酬獲得に貢献するルールの価値は推定値に収束し，貢献しないルールの価値は減少する．一方，式(2)で更新する Profit Sharing では，エピソード内で選択したルールに対し価値を累積するため，収束せず発散する．そのためこれら両手法での価値の大きさは異なった意味を持つ．この観点から，行動選択方法について説明する．

強化学習では十分な学習を行うまでは，それまでの時系列で蓄積した価値を用いて行動を選択する知識利用の方法と，新しい価値を見つけるためにそれまでの知識と異なる観点でランダムに行動を選択する探索行動が必要となる．ε グリーディ手法では確率εでランダム選択を行い，確率 1-ε で最大の価値を持つルールを選択する．Q-learning では学習の収束に伴って最大の Q 値を持つルールが最適解となるため，確率ε を時間に応じて 0 に近づける．一方，Profit Sharing では最適解の保証がないため，最大の価値を持つルールの選択が適しているとは限らない．また，各ルールの価値は発散するため，ε グリーディ手法のような価値の大小関係で比較する方法は適していない．そのため式(2)で更新する Profit Sharing では，価値の相対関係から確率的に行動を選択するソフトマックス行動選択を用いることが多い．ソフトマックス行動選択では，各ルールの価値の重みに基づいて行動選択を行う．ルールの最大値に近い価値を持つルールが存在するとき，そのルールを選択する可能性が生じる．POMDP 環境において，価値の高いルールと低いルールを誤認する場合があるが，ソフトマックス行動選択を使うと，このよ

うな報酬獲得に対して必要な行動選択が複数存在する場合にも対応できる．

ソフトマックス行動選択の一つであるルーレット選択では n 個の行動群の中からあるルール(s, a)の選択確率π(s, a)を式（5）で決める．

$$\pi(s,a) = \frac{Q(s,a)}{\sum_{b=0}^{n-1} Q(s,b)} \quad \cdots(5)$$

また Gibbs 分布に基づいたボルツマン選択がある．ボルツマン選択では温度係数 T を設定する．

$$\pi(s,a) = \frac{e^{Q(s,a)/T}}{\sum_{b=0}^{n-1} e^{Q(s,b)/T}} \quad \cdots(6)$$

温度係数 T を 0 に近づけるとグリーディ手法と同じ選択方法となり，T を無限大に近づけると全てのルールが同じ選択確率　すなわちランダム選択となる．

3　マルチエージェント

同一の問題環境内に複数のエージェントが存在しており，それぞれが学習を行う状況をマルチエージェント環境と呼ぶ（Fig. 1）．マルチエージェント環境下ではエージェト間で相互作用の影響が発生する．各エージェントが協調的な動作を学習できれば 単独のエージェントより複雑な問題を解決したり素早く解決したりできる．一方，マルチエージェント環境下では一般に状態空間が増大するため問題環境全体を観測できず，不完全知覚問題が発生しやすい．特に，エージェント間で情報の共有ができない場合，他のエージェントの内部状態を観測できないため，不完全知覚が発生し，相互作用によって同時学習問題が生じたりする．

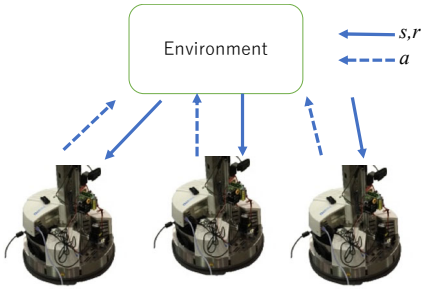

Figure1. Multi-agent

3.1 追跡問題

マルチエージェントの性能評価として追跡問題が用いられることが多い．追跡問題とは任意の $n \times n$ の格子状空間にハンターの役割を持つ複数の追跡エージェントとハンターから逃走する逃走者が存在する問題環境である．ハンター全てが逃走者に隣接する等の特定の状態を捕獲達成として目標状態にすることが多い．ハンターは知覚能力が制限されていることが多く，例えば周囲 m マスのみを観測する．追跡問題では捕獲に際してハンター間の協調行動が必要となる．また，マルチエージェント環境下では囚人のジレンマと同様，単体のエージェントの合理性とエージェント集団での合理性が一致しない場合がある．例えば，各ハンターが逃走者に向かって直線的に追跡すると，互いの進路が干渉し合う場合がある．このような場合には，追従する動作や進路の譲り合いといった協調行動の学習が必要になる．

4　更新型 EPS

EPS を含む価値を累積する累積型 Profit Sharing では，報酬獲得に貢献したルールを優先して選択するため，素早く学習できる反面，局所解に陥る可能性がある．一方，Q-Learning のような更新型の強化方法では，Q 値の増加だけでなく減少にも対応するため，局所解から抜け出すことができる．本研究では EPS の強化関数を更新型に拡張した更新型 EPS を提案する．更新型 EPS では更新式に対し学習率α を導入することで，従来の累積型強化から割り当て報酬に向けて価値を更新する推定値型の強化を行う．提案する更新式を式（6）に示す．

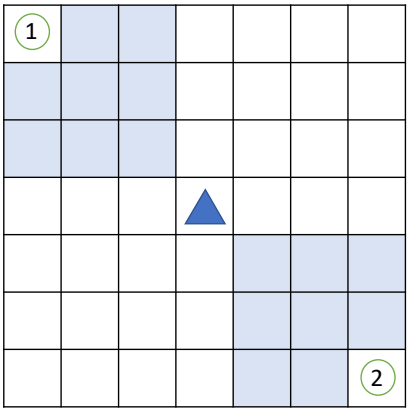

Figure2. The Environment for the Pursuit Problem

$$\omega(s_x, a_x) \leftarrow \omega(s_x, a_x) + r \times f(x)$$
$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\omega(s_t, a_t) \quad \cdots(6)$$

更新型 EPS を用いると，獲得した報酬に対する価値の増減がα によって緩やかになり，環境内の探索が促進される．そのため Profit Sharing において学習に影響を与える Q 値の初期値と獲得報酬値の大小問題[6]の回避が期待できる．

5　実験と結果

ここでは更新型 EPS の効果を確認するため追跡問題の実験を行う．環境は 5×5 マスの格子状空間であり，周囲に壁が存在する（Fig. 2）．壁の通り抜けはできない．また Fig. 3（左）に示す移動式ロボットを追跡エージェントとし，現実世界のロボットによる不審者追跡を想定して環境の知覚能力は逃走者の方が優れているものとする．環境は Fig. 3（右上）のように，追跡エージェント①と②の視界内に逃走者▲が存在するが，追跡エージェント同士はお互いが視界外に存在する状態から始まるケース 1 と，追跡エージェント①の視界内に逃走者▲と味方追跡エージェント②がいるが，追跡エージェント②の視界内に逃走者が存在しない状態から始まるケース 2（Fig. 3 右下）に対してシミュレート実験を行う．結果が Fig. 4 である．

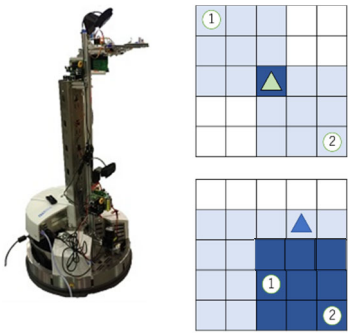

Figure3. Experimental Environment

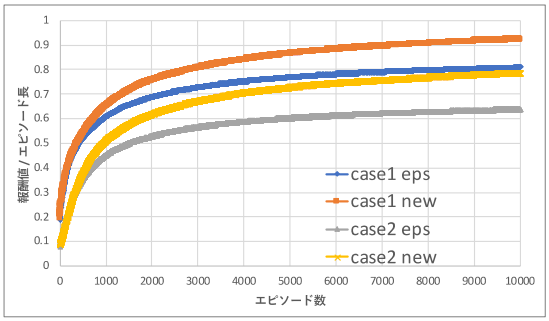

Figure4. Results of Pursuit Problems

6　まとめ

更新型 EPS では EPS による価値の均等分配を行いつつ，更新型強化による価値の急激な増大を抑制することができた．従来の累積型の Profit Sharing は学習の立ち上がり速度が優れている反面，局所解や準最適解を学習する場合がある．そこで価値の割り当て方法として累積型ではなく更新型にすることで，環境の探索性を高めた．

今後は実ロボットへの実装が必要であり，それに伴う問題環境の拡大が必要である．

参考文献

1) Watkins, C. J. C. H. and Dayan, P.：“Technical note:QLearning”, Machine Learning, Vol. 8, pp. 279-292, 1992
2) Whitehead, S. D. and Balland, D. H.： “Active perception and reinforcement learning”, The Seventh International Conference on Machine Learning (ICML ’90), pp. 162 - 169, 1990.
3) Grefenstette J. J. “Credit assignment in rule discovery systems based on genetic algorithms”, Machine Learning 3, pp225-245, 1988
4) 宮崎 和光, 山村 雅幸, 小林 重信:“強化学習にお ける報酬割当ての理論的考察”, 人工知能誌, Vol. 9, No. 4, pp. 580－587, 1994.
5) 植村 渉, 上野 敦志, 辰巳 昭治: “POMDPs 環境のためのエピソード強化型強化学習法”, 信学会論文誌. A, Vol. 88, No. 6, pp. 761-774, 200
6) 植村 渉, 上野 敦志, 辰巳 昭治, : “経験に固執しない Profit Sharing 法”, 人工知能学会論文誌, Vol. 21, pp. 81-93, 2006.