3F4-01

Profit Sharing 法における強化関数に関する一考察

The consideration of the reinforcement function for Profit Sharing

植村 渉　　辰巳 昭治　　北村 泰彦*
Wataru Uemura　　Shoji Tatsumi　　Yasuhiko Kitamura

大阪市立大学工学研究科電子情報系専攻
Dept. of Physical Electronics and Informatics, Graduate School of Engineering, Osaka City University

In this paper, we discuss profit sharing that is one of the reinforcement learning methods. On this method, the agent distributes the received reward to each learning value on the episode(the sequence of selected actions) with reinforcement function of the distance to the goal when it reaches the goal. We propose the condition of a reinforcement function to learn efficiently. The agent should reinforce the action that escapes from a loop, when the episode has one or more loops. The sequence of actions between each loop is noted a zone, and the agent uses the constant value as a reinforcement function at each zone. The value of each zone consists of values of zone reinforcement function. Using this function, the learning result has no loop. And this function can take constant value if the episode has no loop. So we can realize that the agent can learn efficiently without reference to the length of the episode.

1. はじめに

機械学習の一手法として試行錯誤を基にした強化学習がある．エージェントが問題環境との相互作用から解を見つけるため，設計者には問題環境の予備知識を必要としない特徴がある．

強化学習では，環境の状態の表現 $s_t \in S$(S は可能な状態の集合)をエージェントは受け取り，その状態で実行可能な行動の集合から，学習結果に従ったルールを用いて一つの行動を選び，出力とする．そしてその行動の実行により，状態が変化する．目標とする状態に遷移したときのみ，報酬 r が与えられる．強化学習の目標は，この獲得する報酬の総和を素早く大きくすることにより所定の目的を達成することである．

ここで，強化学習独特のトレードオフがある．より良い報酬獲得のためには，環境の同定が必要であるが，その間は報酬の獲得量は期待できないため，環境同定と報酬獲得のどちらを優先するかというトレードオフである．環境同定型の強化学習はその性質上，環境がマルコフ性を満たしていないと効果が現れない．報酬獲得を優先する強化学習は，報酬を獲得できる方法をその都度学習するので，環境の緩やかな変化にも対応できる．本研究で対象とする Profit Sharing 法[Grefenstette88]は報酬獲得を優先する経験強化型の強化学習である．

Profit Sharing では報酬を獲得するまでの行動系列をエピソードと呼び，エピソード単位で強化を行う．獲得した報酬 r をエピソード内の各行動選択の評価値に分配することで強化を行う．この分配関数を強化関数 f と言う．従来の研究では強化関数の設定方法は場当たり的であったが，1994 年に強化関数の定理「合理性定理[宮崎 94]」が提案されて以来，その定理に従った強化関数を用いるようになった．しかしこの定理は，0 への収束が速い関数(代表的な関数として等比減少関数)を要求する為，エピソードが長いと学習速度が極端に遅くなる傾向がある．本研究では，学習の効率を高めた強化関数の条件を提案する．

連絡先：植村 渉，大阪市立大学大学院工学研究科電子情報系専攻情報通信工学講座知識情報処理工学研究室，〒558-8585 大阪市住吉区杉本 3-3-138, Tel & Fax:06-6605-2778, wataru@kdel.info.eng.osaka-cu.ac.jp
*: 現在 関西学院大学理工学部情報科学科

以下第 2 章で合理性定理について紹介し，第 3 章では無駄な学習をしないための効率的な強化関数の条件を提案する．第 4 章では迷路を用いたシミュレーション実験を行い，第 5 章でまとめとする．

2. 従来の手法

あるエピソードにおいて同一状態が二回以上存在し，それぞれで別の行動を選択しているとき，状態遷移を考えるとループが存在する．この時，ループへの行動選択よりもゴールへの行動選択を強化すべきである．このようなループを迂回系列と呼ぶとき，常に迂回系列上にあるルールを無効ルール それ以外のルールを有効ルールという．合理性定理は，エピソード内に無効ルールと有効ルールが存在するときに必ず有効ルールを強化するための条件式(1)を与えている．

$$L\sum_{j=i}^{W} f_i < f_{i-1} \qquad \forall i = 1,2,\ldots,W. \qquad (1)$$

ここで，W はエピソードにある行動数，L は各行動で選択できる行動数の最大である．

また Profit Sharing では報酬を得ることで学習を行うため，この強化関数 f を用いて学習した結果，報酬へたどり着く必要がある．無限にルールを選択し続けるものをプランとし，単位行動当りの報酬の期待値が 0 でないプランを報酬プランとする時，式(1)を満たす強化関数が報酬プランを学習できることが証明されている．

3. 提案手法

合理性定理は無効ルールの抑制を保障しているが，迂回系列中のルールは，迂回系列を抜けるための強化が必要であり，抑制の対象とする必要がない．本研究では迂回系列へ至るルールのみを抑制の対象とする．

エピソード内で，同一状態が複数存在し，異なるルールを選択した時，それらのルールに対して目標状態に一番近いルールを非迂回ルール，それ以外のルールを迂回ルールとする．非迂回ルールや迂回ルールを区切とし，エピソードのルール群を分割する．それぞれの分割したルール群を区間 Z とし，目標状態に近い順に z_1, z_2, …, z_n とする．一般的に区間内のルール群に優越はないので区間ごとに同一の強化値で強化する．区

間に対する強化関数として区間強化関数 g を用いる．区間内の強化に減少関数を用いても問題はないが，次の区間の強化値を下回ってはいけない．ここでエピソードの i 番目のルールが有効ルールであり，i+1 番目のルールが無効ルールの時，この i+1 番目のルールは迂回ルールとなる．つまり，無効ルールが一つ以上連続するとき，先頭の無効ルールは必ず迂回ルールとなる．この時，次式(2)を満たす区間強化関数 g が迂回ルールを抑制できることになる．

$$L\sum_{j=i}^{n} g_i < g_{i-1} \qquad \forall i = 1,2,\ldots,n. \qquad (2)$$

ここで，n は区間の数である．

以下に，このことを説明する．

3.1 迂回ルールの抑制

迂回系列への遷移があるエピソードを考える．迂回系列内にあるルールの数を n とし，非迂回ルールを用いて迂回系列から抜け出し，目標状態に至るまでのルールの数をω，区間数を z とする．この時，迂回系列を抜け出る行動(非迂回ルール)の学習には f_{ω}を用い，迂回系列内の行動の学習には $f_{\omega+1}$，…，$f_{\omega+n}$ を用いる．

ここで，迂回ルールが学習時に一番強化されるのは，迂回系列内の行動強化に用いる強化値全てを用いて強化する時で，値は $\Sigma_{i=\omega+1}^{\omega+n} f_i$ である．この時，非迂回系列を g_{z-1} で強化し，迂回ルールを $\Sigma_{i=z}^{z+n} g_i$ で強化する．この状況で迂回ルールを抑制する必要がある．よって，次式(3)を満たす必要がある．

$$\sum_{i=z}^{z+n} g_i < g_{z-1} \qquad (3)$$

ここで，非迂回ルールの数 L が複数の時を考える．一番選択確率の高い非迂回ルール A が，全非迂回ルールの中から選択される確率は 1/L 以上である．この非迂回ルール A が選択された後，最悪 L 回他の非迂回ルールが選択される場合を考える．迂回ルールを強化した強化量よりも非迂回ルール A の強化量が大きい必要がある．以上をまとめると次の定理を得る．

[定理 1]迂回ルールの抑制

任意の迂回ルールが抑制される必要十分条件は，下記の不等式(4)が成立することである．

$$L\sum_{j=i}^{n} g_i < g_{i-1} \qquad \forall i = 1,2,\ldots,n. \qquad (4)$$

ここで，n は区間の数，L は迂回ルールの状態にある非迂回ルールの数である．L はその状態で選択できる行動数とすることで十分である．以後式(4)を迂回ルール抑制条件と呼ぶ．

3.2 報酬プランの獲得

迂回ルール抑制条件を満たす強化関数を用いて学習を行った際，報酬を得られないプランに陥るかどうか検討する．

[定理 2]報酬プランの獲得

強化関数が迂回ルール抑制条件を満たせば，報酬プランを必ず獲得できる．

(証明は付録 A)

3.3 学習に必要とする時間

迂回ルール抑制条件を満たす強化関数を用いて，学習する時の学習時間と環境の大きさの関係について考える．条件を満たす関数を用いて学習すると，その学習対象となる行動系列ではループが抑制される．行動系列にループがない場合は区間数が 1 であるため，定数を用いて強化することが許され，強化関数に環境の大きさが含まれない．Profit Sharing は目標状態に到着して初めて学習が行われるため，学習の立ち上がりの時間は学習初期のランダムウォークに起因する．

1 次元の壁なし迷路の場合，ゴールに到達する確率が 50%となる試行回数は，ゴールまでのステップ数の倍を必要とし，ゴールまでの距離に比例する．多次元の場合，ゴールと同じ軸上以外ではゴールに近づく行動を解とすると，解と解以外の行動(ゴールから遠ざかる行動)の割合が 1 次元と同じになるため，試行回数はゴールまでの距離にほぼ比例する．状態での解の割合が 1/2 より少なくなると試行回数は多くなる．つまり，学習時間は環境の解の割合に影響し，解の割合が 1/2 の場合は環境の大きさに比例した時間がかかる．例えば，先手後手の優越がないボードゲームは勝率 1/2 のため，環境の大きさに比例した時間での学習が期待できる．

3.4 従来の手法との比較

従来の合理性定理と本定理の関数クラスの範囲を図 1(a)に示す．本定理は抑制する対象を限定したため，扱える関数の種類が多くなっている．また抑制対象を図 1(b)に示す．有効ルールの存在しない状態での無効ルールの抑制動作が異なる．有効ルールの存在しない状態とは，常に迂回系列上にある状態である．しかし，その状態を起点と考えると，迂回系列から抜け出るためのルールが存在し，そのルールは学習すべきであり抑制する対象ではない．つまり，本定理は迂回系列に陥っても回復するための学習ができ，学習効率がよいことがわかる．

4. 実験と結果

迂回ルール抑制条件に従った強化関数を用いた学習の効果を，迷路を用いたシミュレーション実験にて確認する．実験環境は図 2 に示す迷路走行タスク[Sutton 98]を用いた．始点(S)から終点(G)までの経路を学習する問題である．乱数系列を変えた実験を 100 回行い，その平均値を実験値とする．最適解の値は，報酬までの最短経路 14 ステップより，10/14　0.714 である．選択できる行動数 S=4 のため，Profiot Sharing の強化関数

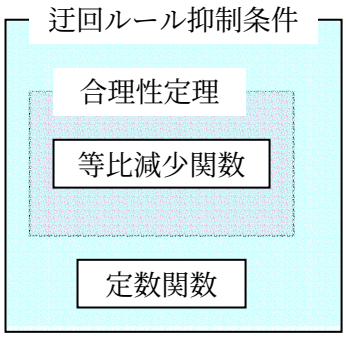
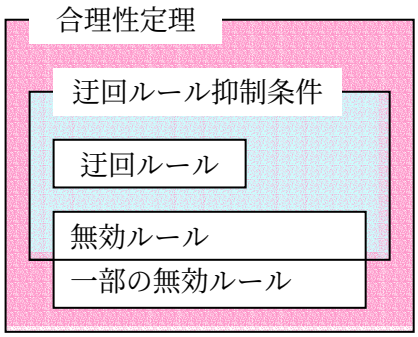


図 1 (a) 定理に従う関数の範囲



(b) 抑制対象

5	11	14	20	26	31	37		G
4	10		19	25	30	36		45
S	9		18	24	29	35		44
3	8		17	23	28	34	40	43
2	7	13	16	22		33	39	42
1	6	12	15	21	27	32	38	41

図 2 実験に用いた迷路環境

は公比 1/4 の等比減少関数を用い，提案手法の区間強化関数も公比 1/4 の等比減少関数を用いた．

また，環境を複雑にした時の性能比較のため，迷路の縦横の長さをそれぞれ 2，3 倍にした迷路の実験を行った．

結果は図 3，図 4 である．従来の手法では，強化関数として常に減少する関数を用いたが，本提案手法では必要なときだけ減少する関数となるため，その改善効果が学習速度の違いとしてあらわれることが確認できる．

環境を複雑にすると従来の手法では学習の効果が確認できないが，本手法では効果を確認できる．収束値の 90%の値に達する行動選択回数は，二倍迷路の時で約 22000 回，三倍迷路で約 45600 回である．状態数の増加に対して収束までの時間が線形的な増加量を示していることが図 5 でわかる．迂回ルール抑制条件を満たす強化関数を用いて迷路走行タスクを学習すると，学習時間は環境の大きさに比例し，安定して学習できることがわかる．

5. 考察とまとめ

強化学習 Profit Sharing を用いて学習する際，従来は無効ルールを抑制するために等比減少関数を強化関数に用いた．しかし，強化関数の 0 への収束速度が速いため，環境が複雑になると学習効率が悪くなることがしばしばであった．

本研究では，抑制する対象を無効ルールの先頭である迂回ルールのみに絞る方法を提案した．報酬プラン獲得の条件を満たしながら，強化関数の 0 への収束速度を改善でき，環境の複雑さに影響を受けにくい学習が実現した．

本研究により環境の複雑さの制限がなくなり，今まで状態数が大きすぎて Profit Sharing を導入できなかった問題への適用が今後期待される．

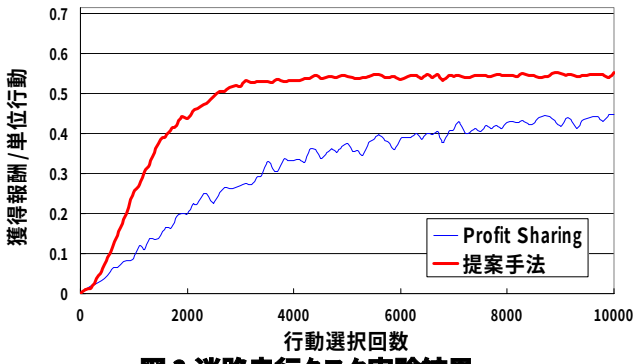


図 3 迷路走行タスク実験結果

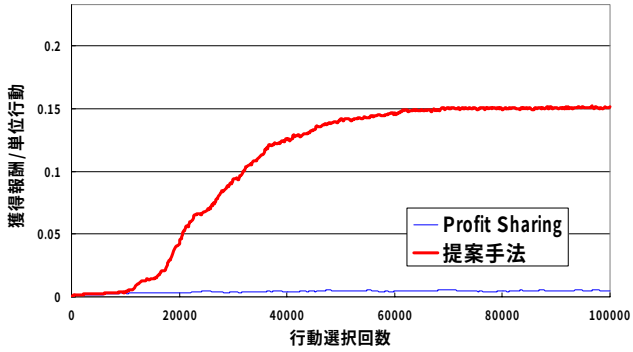


図 4 縦横三倍の迷路における実験結果

参考文献

[Grefenstette 88] Grefenstette, J.J., “Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms,” Machine Learning, Vol.3, pp.225-245(1988).

[宮崎 94] 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).

[Sutton 98] Richard S.Sutton and Andrew G.barto, “Reinforcement Learning”, The MIT Press.(1988)

A. 付録:定理 2 の証明

[宮崎 94]の証明と同様の環境を用いる．迂回ルール抑制条件を満足する強化関数を用いて学習し，その学習結果に従って行動した時，報酬プランとならない場合を考える．報酬プランとならない場合は，報酬を伴わないループに陥る時である．このようなループは 2 個以上のエピソードから構成される．ここで有効ルール数 L=2 かつループから脱出できる行動(x_o,y_o)のある状態を x, y の二ヵ所とする(図 6 参照)が，それ以外の場合もまったく同様である．状態 x, y 間でループを構成するためにはループの出口となる状態において次の不等式群が成り立つ必要がある．

$$\Delta\omega\vec{x}_o<\Delta\omega\vec{x}_i \quad (A.1)$$

$$\Delta\omega\vec{y}_o<\Delta\omega\vec{y}_i \quad (A.2)$$

$\vec{x}_o$, $\vec{y}_o$はそのループから出て行くルール，$\vec{x}_i$, $\vec{y}_i$はそのループ内に戻るルールである．また，Δはそのルールに加算される強化値の総和を表す．迂回ルール抑制条件を満足し，かつ$\vec{x}_i$を含むエピソードが$\vec{x}_o$を含んでいたとすると，

$$\Delta\omega\vec{x}_o>\Delta\omega\vec{x}_i \quad (A.3)$$

となり，ループが構成できない．よって，$\vec{x}_i$を含むエピソードは$\vec{x}_o$以外のルール つまり$\vec{y}_o$を使ってループの外へ出る必要がある．$\vec{y}_i$に関しても同様である．次の不等式群が成り立つ．

$$\Delta\omega\vec{x}_i\leqq\Delta\omega\vec{y}_o \quad (A.4)$$

$$\Delta\omega\vec{y}_i\leqq\Delta\omega\vec{x}_o \quad (A.5)$$

等号成立はそれぞれ同一区間にある場合

式(A.1)，式(A.2)，式(A.4)と式(A.5)より，次の不等式が得られる．

$$\Delta\omega\vec{x}_o+\Delta\omega\vec{y}_i\leqq\Delta\omega\vec{x}_o+\Delta\omega\vec{y}_o<\Delta\omega\vec{x}_o+\Delta\omega\vec{y}_i \quad (A.6)$$

この不等式を満たす解は存在しない．ゆえに，迂回ルール抑制条件を満たす区間強化関数を用いた強化関数では，必ず報酬プランが獲得される．

(証明終了)

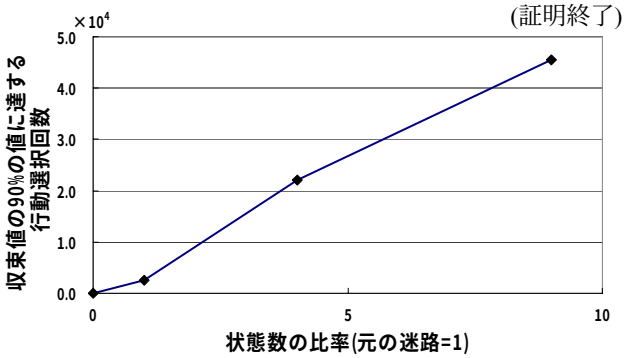


図 5 収束速度と状態数の関係

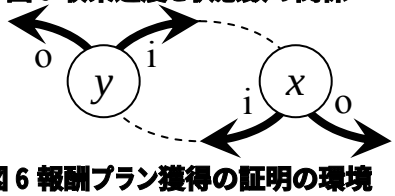


図 6 報酬プラン獲得の証明の環境