強化学習における探査率の動的制御

室蘭工業大学 情報電子工学系専攻 認知ロボティクス研究室 澁谷和

**1. はじめに**

　近年，実ロボットに適用される手法として強化学習[1]が注目を集めている．強化学習は，ある状態で取った行動の結果に着目し，このときの評価が良くなるように学習を行うものである．

このときに利用するのが報酬と呼ばれるスカラ値の情報である．ロボットは行動を取ることでその行動に見合った報酬が得られる．人間は報酬さえ設定すれば，あとはロボットに任せればロボットは報酬獲得までの行動を自動的に獲得できる．

　強化学習の問題点の一つに探査-利用のトレードオフ問題がある．多くの報酬を獲得し続けるには，過去に試みた行動の中で，最も多くの報酬を獲得できるような行動を取り続けなければならない．このような行動を取り続けることは，ロボットが所有している知識を利用していることとなる．しかし，現在所有している知識が最適なものとは限らない．そのため，より多い報酬を獲得するためには過去に試みていない行動を行わなければいけいない．つまり，未知の状態を経験することが必要である．このような未知の状態を経験するために行動することを探査という．探査と利用の行動のうち，たとえ同じ状態でもその状態に対して，探査か利用かを取るべきかということは一意に定まらない．そのため，ロボットが探索と利用のバランスを自律的に制御することが必要である．

　強化学習において，探索と利用のバランスは行動決定法のパラメータによって決定する．本論文ではε-greedy法を対象にする．ε-greedy手法は，現在の行動価値が最も高い行動を(1-ε)の確率で選択(利用)するか，探査率εでランダムに行動を選択(探査)するという手法である．εが小さいほど，現時点で最適とされる行動が行われる回数が多くなるが，真に最適な行動を見つけ出すまでに時間がかかってしまう．一方，εが大きいほど，最適な行動を見つけやすくなるが，発見した後もランダム行動を取り続けるため，無駄な行動が多くなる．それゆえ，εの値を調整し，探査と利用のバランスの取り方を考える必要がある．

　従来手法としては，εを学習初期で高くし，学習が進むにつれて徐々に低くするという方法が挙げられる．しかし，この方法は静的環境では有効であるが，環境が変化する動的環境に対しては対応が難しい．また，遺伝的アルゴリズムでεを調整する研究[2]もあるが，こちらも同じく動的な環境には向かない．

そこで，本論文では，環境変化に対応できるようにロボット自身が直面する状況に応じて，探査率εを自律的に制御するシステムの構築を目指す．

**2. 情報量を用いた探査率の制御**

本論文では遷移先の情報量に応じて，探査率εを調整することを提案する．ある状態である行動を取った時の遷移先が一意に決まる場合は現在の知識を用いて，利用行動にしたほうが良い．また，遷移先が確定しないもしくは未経験である場合には探査をしたほうが良く，εは高くするのが理想である．そこで，遷移先が確定するかしないかという指標を情報量で表す．

本研究では提案手法は2つのモジュールから構成される(図1)．1つは「経験情報の獲得」である．経験情報の獲得では，環境のモデルを知るために，エージェントは行動回数と遷移回数を行動毎に記録する．その情報からある状態のときある行動を取ると，どのくらいの確率で次状態に遷移するかという遷移確率を算出する．遷移確率の算出は式(1)で行う．

$P\_{a\_{j}}\left(s\_{i},a\_{j}\right)=\frac{R(s\_{i},a\_{j},s\_{k})}{N\left(s\_{i},a\_{j}\right)+R(s\_{i},a\_{j},s\_{u})}$ (1)

ここで，$N\left(s\_{i},a\_{j}\right)$は状態$s\_{i}$のとき，行動$a\_{j}$を選択した回数である．また，$R(s\_{i},a\_{j},s\_{k})$は状態$s\_{i}$のとき，行動$a\_{j}$を選択したとき，次状態$s\_{k}$に遷移した回数である．

2つめのモジュールは「探査率εの算出」である．経験情報の獲得で算出した遷移確率を用いて，各状態行動対に対して遷移先が確定しているのか，未確定なのかを平均情報量で判断する．平均情報量の算出式を式(2)，(3)に示す．式(2)は行動回数が0の場合，つまり未経験の場合であり，式(3)は行動回数が1以上の場合の算出式である．

$H\left(s\_{i},a\_{j}\right)=log\_{2}n$ (2)

$H\left(s\_{i},a\_{j}\right)=-\sum\_{k=0}^{n}P\_{a\_{j}}(s\_{i},s\_{k})log\_{2}P\_{a\_{j}}(s\_{i},s\_{k})$ (3)

ここで，$H\left(s\_{i},a\_{j}\right)$は状態$s\_{i}$における，行動$a\_{j}$を選択した際の，次状態における平均情報量を表す．

そして，その平均情報量を0から1に正規化し，全状態行動対の情報量を合計した値をεとした．εは，試行毎に算出している．εの算出式を式(4)で示す．

$ε=\frac{1}{nm}\sum\_{i=0}^{n}\sum\_{j=0}^{m}\frac{H(s\_{i},a\_{j})}{log\_{2}n}$ (4)



図1：提案手法の概要

**3. 経路計画問題における検証実験**

　提案手法の有効性はシミュレーションを通して検証する．本実験ではεを学習中一定に固定したε-greedy法と提案手法の学習過程を比較する．本実験では経路計画問題を適用し，ゴールに到達するまでの行動数を比較する．ただし，環境変化が起こる．具体的には一定試行ごとに障害物が環境内に発生する．障害物が置かれたマスにはロボットは入ることができない．

　表1に実験で用いたパラメータを示す．また，実験で用いた環境を図2に示す．この環境は障害物を置く前の環境である．

表1：パラメータ設定

|  |  |
| --- | --- |
| 報酬(ゴールのみ) | 100 |
| 総試行数 | 1000 |
| 環境が変化する試行数 | 200 |
| 障害物の数 | 20 |
| 総実験数 | 100 |
| Q値の初期値 | 0.00 |
| 学習率α | 0.1 |
| 割引率γ | 0.9 |



図2：実験に使用した環境

　図3は提案手法におけるεの推移を表す．また，図4に各εと提案手法の累積ステップ数を示す．

　実験結果から，障害物が発生した200，400試行目ではεが上昇し，環境変化に対して，探査行動を増やそうとしている．その結果，環境変化に素早く対応することができている．εが低い場合では，環境変化が発生しても，環境が変化する前の知識を利用し，行動を決定することが多いため，障害物を避ける事ができない．結果，障害物にぶつかることが多く，環境変化が起きた200,400,800試行付近で，急激にステップ数が増加してしまう．

　しかし，提案手法ではε=0.09やε=0.10よりもステップ数が多くなっている．これは障害発生時にεが0.10程度しか上昇しないためである．原因としては，環境変化により情報量は上昇したが，環境の一部のみ変化したため，情報量が変化しない状態がある．そのため，全ての状態行動対に関して，総計した場合，環境変化した部分の影響が弱くなり，εの上昇率が悪くなる．



1. 全体 (b)拡大(Y軸0-0.2)

図3：提案手法におけるεの推移



図4：試行数における累積ステップ数

**4. おわりに**

　本研究では，強化学習における探査率εを情報量によって制御することを提案した．検証実験として，経路計画問題に適用し，ステップ数を比較し，手法の有効性を確認した

　参考文献

[1] Richard S. Sutton and Andrew G. Brato，"Reinforcement Learning"，The MIT Press,　1998

[2] 亀井圭史，石川眞澄，”遺伝的アルゴリズムによる移動ロボットの強化学習パラメータ最適化”，電子情報通信学会技術研究報告. NC, ニューロコンピューティング，2005