行動確率場モデルに基づく強化学習 －拡張Q-学習－

榎田 修一　大橋 健　吉田 隆一　江島 俊朗
(九州工業大学情報工学部)

概要 自律型ロボットの学習による行動獲得は，先見的な知識だけでは補いきれない行動決定の問題に対して有効な手法である．従来，センサ空間を離散化し，有限個の状態上での行動決定問題として定式化され，Q-学習など興味深い学習法が提案されて来た．しかし，離散化に伴う誤差が無視できない状況も多く，そのため誤差の影響を少なくする高精度の方法が研究対象になってきた．

本論文では、Q-学習を拡張した拡張Q-学習を提案する．拡張Q-学習とは，行動確率場モデルに基づき，センサ空間から行動空間への写像を導くものである．本モデルでは写像を表す行動選択確率を規定する行動価値関数は，有限個の基底関数の重み付き和ととして表される．学習は重みを調整する作業に対応し，また，精度を保持しつつより簡潔なモデルで関数近似を行うために基底関数の自律統合を学習アルゴリズムに追加した．

Reinforcement Learning Based on Stochastic Field Model –Extended Q-Learning–

Shuichi Enokida　Ohashi Takeshi　Takaichi Yoshida　Toshiaki Ejima
(Kyushu Institute of Technology)

Abstract: Reinforcement learning has been used as a method that makes an autonomous robot to select an appropriate action in each state through an interaction with an environment. Typically, even if the autonomous robot has continuous sensor values, sensor space is quantized to reduce learning time. However, the reinforcement learning algorithms including Q-learning suffer from errors due to state space sampling. To overcome the above, we propose Extended Q-learning (EQ-learning) based on Q-learning which creates mapping that maps a continuous sensor space to a descrete action space. Through EQ-learning, action-value function approximation is represented by a summation of weighted base functions, and the autonomous robot adjusts only weights of base functions by robot learning. Other parameters are calculated automatically by unification of two similar base functions.

1 はじめに

自律型ロボットとは，センサで環境を認識し，適切な行動を自ら決定し，環境に動作するものである．環境との相互作用によりロボットがとるべき行動を自律的に獲得する方法として，強化学習がある．カメラ画像等の連続量のセンサ空間を持つロボットの学習では，センサ空間をそのまま状態空間と考えると状態が膨大な数になり，学習モデルとしては現実的でない．そこで，ロボットのセンサ空間を適切に離散化した状態空間の構築が必要であるが，センサ空間を離散化することにより誤差が生じる．結果，学習により得られる行動も誤差を含んだものとなる．そのため，誤差の影響が少なく学習効率の良い手法の研究が盛んになってきた．

本論文ではセンサ空間の離散化を行わずに，行動確率場モデルによりセンサ空間からK次元確率ベクトル空間への写像を求める手法を提案する．ここでのK次元確率ベクトル空間とは，ロボットが取り得るK個の行動それぞれに対する選択確率である．行動選択確率は連続量の行動価値関数により導かれるとする．行動価値関数を基底関数の重み付き和で表し，学習により簡潔で近似精度のよい基底関数の組合せを導き出すことを考える．

提案する学習アルゴリズムでは，基底関数として局所性をもつ方形波関数とガウス関数に注目する．方形波関数を用いると，従来の離散状態での強化学習法と等価となり，ガウス関数を用いると，連続系の学習となる．ガウス関数を用いることにより，離散化した状態での学習の際に発生していた誤差を吸収し，より高精度の学習が可能となる．

モデルの善し悪しを測る手法としてAIC[3]がある

が，その評価基準はモデルと観測される実際の値との誤差，モデルの簡潔さの二点を持って与えられる．そこで，基底関数の類似に注目し，類似するもの同士を統合することにより，学習データを反映した適切な簡潔さを有するモデルを自律獲得することを目指す．

2 Q-学習

強化学習法として広く用いられている，Watkinsにより提案されたQ-学習 [1] について述べる．ロボットが，状態 $s \in S$ で行動 $a_i \in A$ を取り，次状態 s' に遷移したときに，報酬 $r(s,a_i)$ を受けたとする．そのとき，報酬を手がかりに，行動価値 $Q(s,a_i)$ を以下の式で更新し，報酬の得られた行動の価値を高め，その行動を選ぶ機会を強化する学習法である．

$$Q(s,a_i) \Leftarrow Q(s,a_i) + \alpha(r(s,a_i) + \gamma M(s') - Q(s,a_i)) \quad (1)$$

$$M(s) = \max_{j \in A} Q(s,a_j) \quad (2)$$

ここで，α は学習率，γ は減衰係数である．学習時の報酬は，一連の動作結果で仕事をうまく遂行できたときにのみ入力されるため，目的状態に到達する1ステップのみの行動価値が強化される．そこで，行動価値の時間的差分を最小にすることにより行動価値が目的状態から徐々に伝播し，各状態での行動価値が得られる．また，行動価値に基づき以下の式で行動を選択する確率を求める．

$$p(\boldsymbol{x},a_i) = \frac{\exp(U(\boldsymbol{x},a_i)/T)}{\sum_{j=1}^{K} \exp(U(\boldsymbol{x},a_j)/T)} \quad (3)$$

$$\boldsymbol{U}(\boldsymbol{x}) = (U(\boldsymbol{x},a_1), U(\boldsymbol{x},a_2), \cdots, U(\boldsymbol{x},a_K)) \quad (4)$$

ここで T は温度定数である．

連続系のセンサ状態に対して，強化学習を適用することを考える．強化学習によりセンサ空間全てに対し行動価値を決定することは，学習時間の爆発的な増大をまねき，非常に困難であり現実性に欠ける．そこで，一般的にQ-学習を用いるときは，センサ空間を離散化しカテゴライズする．そして，学習のときに報酬が得られると同一状態に対して一律に行動価値の更新を行うことにより，学習時間の爆発を抑えている．

3 拡張Q-学習

本論文では行動確率場モデルによる拡張Q-学習（以降EQ-学習）を提案する．本モデルではセンサ空間から K 次元の確率ベクトルへの写像を学習により導く．行動選択確率 $p(\boldsymbol{x},a_i)$ は式（3）により決定される．また，式（3）での行動価値関数をセンサ空間で近似する際に，基底関数の重み付き線形和を用いる．

3.1 連続状態での行動価値関数

状態 $\boldsymbol{x}$ における行動 a_i の行動価値 $U(\boldsymbol{x},a_i)$ を N 個の基底関数 $B_m(\boldsymbol{x})$ の重みつき線形和で表す．

$$U(\boldsymbol{x},a_i) = \sum_{m=1}^{N} W_m(a_i) B_m(\boldsymbol{x}) \quad (5)$$

いま，行動価値関数を有限個の基底関数で近似しようとしたとき，モデルを規定するものは，基底関数（ガウス関数，シグモイド関数等），基底関数の自由パラメータ（中心座標，分散等），基底関数の次数，重みである．

離散化したセンサ空間での学習は基底関数 $B_m(\boldsymbol{x})$ を，以下の式で示す方形波関数で表されるものを用いてきたと捉えることができる．EQ-学習は，従来のQ-学習の手法を含む一般的な枠組となるものを目指す．また，本論文では，基底関数 $B_m(\boldsymbol{x})$ をガウス関数で表すものを新たに提案し，従来法との比較を行う．

3.2 EQ-学習における重みの更新 [2]

本論文では，経験を通して強化信号を伝播し，期待利得が最大となるような行動を獲得する強化学習に基づき，重みの更新学習法について考える．特に，各状態で個々の行動を選択したときの期待利得を逐次的に求めてるQ-学習に焦点を当て，Q-学習が行動確率場モデルによりセンサ空間でも働くように，学習アルゴリズムの拡張を図る．

行動確率場モデルでの学習更新式を，最急降下法により導く．センサ空間の次元数に影響されない重みパラメータを以下の式で更新する．

$$W_m(a_i) \Leftarrow W_m(a_i) + N_m(\boldsymbol{x})(r(\boldsymbol{x},a_i) + \gamma M(\boldsymbol{x}') - U(\boldsymbol{x},a_i)) \quad (6)$$

$$N_m(\boldsymbol{x}) = \alpha \frac{B_m(\boldsymbol{x})}{\sum_j B_j(\boldsymbol{x})} \quad (7)$$

$$M(\boldsymbol{x}) = \max_{k \in A} U(\boldsymbol{x}, k) \quad (8)$$

この更新式に従って各重みを更新する．

3.3 基底関数の自律統合

基底関数の絶対数を増加させることにより，行動価値観数の近似精度は良くなることが考えられるが，不必要にパラメータの多いモデルでは学習時間の増大を招く．また，モデルの善し悪しを評価するAICにおいても，同じ近似精度を持つモデルであれば，そのモデルの持つ自由パラメータが少ない程良いと評価される．そこで，EQ-学習では基底関数の自律統合により近似精度を保持しつつ，モデル内の自由パラメータ数の削減を試みる．

3.3.1 行動選択確率の類似度

ある基底関数 $B_n(\boldsymbol{x})$ の中心座標での行動選択確率ベクトルの，$B_m(\boldsymbol{x})$ の中心座標での行動選択確率ベクトルに関するKullback-Leibler情報量は，

$$I(P(\boldsymbol{\mu}_n); P(\boldsymbol{\mu}_m)) = \sum_{i=1}^{K} p(\boldsymbol{\mu}_n, a_i) \log \frac{p(\boldsymbol{\mu}_n, a_i)}{p(\boldsymbol{\mu}_m, a_i)} \quad (9)$$

である．この値が0に近い程，行動選択確率分布の類似度が高いとする．

3.3.2 中心座標の類似度

ある基底関数 $B_n(\boldsymbol{x})$ の中心座標から $B_m(\boldsymbol{x})$ の中心座標までのマハラノビス距離 d_{nm} は，

$$d_{nm}^2 = (\boldsymbol{\mu}_m - \boldsymbol{\mu}_n)^t \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu}_m - \boldsymbol{\mu}_n) \quad (10)$$

となる．マハラノビス距離が0に近い程，中心座標の類似度が高いとする．

3.3.3 基底関数の類似度

行動選択確率と中心座標の類似度より基底関数の類似度とする．今，基底関数 $B_n(\boldsymbol{x})$ の $B_m(\boldsymbol{x})$ に対する類似度 $s(n,m)$ を

$$s(n,m) = \exp(-aI(P(\boldsymbol{\mu}_n); P(\boldsymbol{\mu}_m)) - bd_{nm}) \quad (11)$$

とする．ここで，a, b はともに任意の定数である．類似度の最大値は1であり，類似度が低い程0に近付く．

また，類似度を評価する時には，学習が進み，各基底関数の自己組織化が進むなかで，高い類似度を持ち続ける基底関数こそ高い類似度を持つとすべきある．そこで，類似度 $S_{nm}(t)$ を新たに定義する．

$$S_{nm}(0) = 0 \quad (12)$$

$$S_{nm}(t+1) = (1-\beta) S_{nm}(t) + \beta s(n,m) \quad (13)$$

ここで，β は $0 \leq \beta \leq 1$ の値をとる定数である．また，このときの単位時刻 t は，報酬が入力されたときに進むものであるとする．

3.3.4 基底関数の統合

閾値 th 以上の類似度を持つ基底関数 $B_n(\boldsymbol{x})$ と，$B_m(\boldsymbol{x})$ とを統合する．統合の結果，新たに生成される基底関数 $B_{new}(\boldsymbol{x})$ のパラメータを，

$$\boldsymbol{\mu}_{new} = \frac{d_{mn}}{d_{nm}+d_{mn}} \boldsymbol{\mu}_n + \frac{d_{nm}}{d_{nm}+d_{mn}} \boldsymbol{\mu}_m \quad (14)$$

$$\boldsymbol{\Sigma}_{new} = \boldsymbol{\Sigma}_n + \begin{pmatrix} \Delta x^2 & 0 \\ 0 & \Delta y^2 \end{pmatrix} \quad (15)$$

$$\Delta x = \mu_{new}(x) - \mu_n(x) \quad (16)$$

$$\Delta y = \mu_{new}(y) - \mu_n(y) \quad (17)$$

とする．ここで，任意の重み $W_{new}(a_i)$ は，

$$W_{new}(a_i) = W_n(a_i) B_n(\boldsymbol{x}) + W_m(a_i) B_m(\boldsymbol{x}) \quad (18)$$

とする．これは，新たに置き換えられる基底関数の中心座標における行動価値関数の値を保持するためである．統合の結果生成された新たな基底関数に関する類似度は，すべて0に初期化する．

4 実験

実験対象のロボットは餌までの距離，視界端からの角度の2次元ベクトルがセンサ空間として入力される．また，選択可能な行動は，前進・左右前進・左右旋回の5通りである（図1）．ロボットは餌を拾った時のみ報酬を獲得する．また、学習時の各パラメータは $\alpha = 0.25, \gamma = 0.99, T = 0.1$ とする．

各学習回数（報酬入力）毎に学習を停止し，各有用度関数を用いて求餌行動に適用し，餌を拾うまで

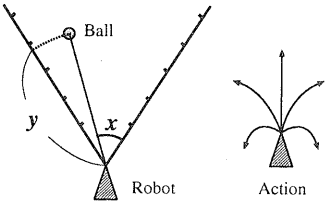


図 1: ロボットパラメータ

にかかったステップ数を評価する．評価は，学習し，個性を持つ 30 台のロボットが，それぞれ環境内に散在する 3 つの餌を拾い終えるまでにかかったステップ数を 1000 回ずつ記録し，その平均ステップ数を測定することにより行う．また，評価を行う際に 10000 ステップ以上必要であったもののステップ数は 10000 とした．

4.1 実験結果

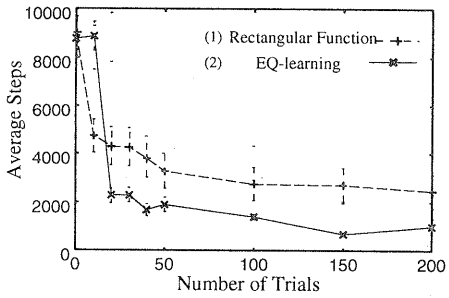
EQ-学習の評価実験を行う．基底関数をセンサ空間内に格子状に 4×4 配置し，初期状態とする．ここで，実験（1）方形波関数，自律統合無し，は従来の Q-学習に対応し，それに対する学習の収束の具合を評価し，行動価値の近似精度の評価とする．図 2 に実験（1）と実験（2）EQ-学習の結果を，自律統合される基底関数の例とともに示す．自律統合の各パラメータは $a = 1, b = 0.5, \beta = 0.1, th = 0.25$ と定めた．

4.2 考察

EQ-学習による学習の収束性の高さ，行動の高精度化が確認された．これは，類似度の高い基底関数同士を統合することにより，学習パラメータが絞られた結果である．また，適切な基底関数を自動獲得し，学習が可能であることが確認された．しかし，統一の際に誤差が大きくなることもあるために，仕事効率が落ちることがある．

5 まとめと課題

本論文では，強化学習の一つとして Q-学習を拡張した EQ-学習を提案した．EQ-学習は，行動価値関数



Number of trials	0	50	100	200
Average numbers of base functions	16	4.2	2.8	2.5
Example of the unification				

図 2: EQ-学習による学習結果と基底関数総数 (上) と基底関数の統合例 (下)

を基底関数の重み付き和であらわすことにより，連続系のセンサ状態をそのまま扱うことが可能である．また，基底関数の自律統合により，近似精度をある程度保ちつつ，より簡潔なモデルを自律獲得可能となった．基底関数の類似に注目し，類似するもの同士を統合することにより実現した．結果，学習を行うときに基底関数の配置をタスク毎に構成する必要がなく，タスク，ロボットパラメータに対して適切，かつ簡潔なモデルを自律的に獲得可能である．

今後の課題として，統合だけの一方通行的な基底関数配置手法に改善の余地がある．統合を行った後に誤差が大きくなり，仕事遂行の効率が著しく落ちたときには再度基底関数を分割するなどの対処が必要である．

参考文献

[1] C.J.C.H. Watkins : Learning from delayed rewards, PhD thesis, University of Cambridge, 1989.

[2] 榎田 修一 : 強化学習による行動獲得の高精度化に関する研究，九州工業大学修士論文，1999.

[3] 坂元 慶行，石黒 真木夫，北川 源四郎 : 情報量統計学，共立出版株式会社，第 4 章，1983.