転移学習を用いた強化学習ロボットの方策選択における認知的経済性の検討

Consideration of Cognitive Economics for Transfer Reinforcement Learning with Policy Selection Method

○学　坂本裕都 (工芸大)　　河野　仁 (工芸大)
温　　文 (東大)　　正　藤井浩光 (千葉工大)
正　鈴木　剛 (電機大)

Yuto SAKAMOTO, Tokyo Polytechnic University, e1619025@st.t-kougei.ac.jp
Hitoshi KONO, Tokyo Polytechnic University
Wen WEN , The University of Tokyo
Hiromitsu FUJII, Chiba Institute of Technology
Tsuyoshi SUZUKI, Tokyo Denki University

This paper is discribes the cognitive economics of the transfer reinforcement learning method using policy selection based on spreading activation model. Proposed policy selection model has some problems: calculation cost is increased by number of reusable learned policies. The computational processing cost increases because of aiming for higher cognitive functions in intelligent robots. As a basic consideration, parallelizing of policy selection method is proposed to increase the speed of processing. In this paper, effectiveness of parallelized policy selection method is confirmed using computer simulation and demonstrated using actual mobile robot.

Key Words: Reinforcement Learning, Transfer Learning, Policy Selection, Autonomous Mobile Robot

1　緒言

近年，社会では少子高齢化などの問題により，人の代わりに自律的に多様なタスクを遂行する知能ロボットが期待されている．知能ロボットには，学習アルゴリズムが有用とされており，身体性を有するロボットに適した学習アルゴリズムとして，自ら試行錯誤的に行動し最適解を学習する強化学習がある [1]．また学習時間短縮のために，強化学習で得た知識を別のタスクに再利用する転移学習という手法が存在する [2]．転移学習はタスクに応じて適した知識選択をしなければ効果が得られない．知識選択において様々な環境やタスクに応じて適した知識を効率的に選択するために SAP-net を用いた転移学習手法が提案されている [3][4]．しかし，実装時に方策選択に要する時間が長いという課題から，効率的な知識選択が実現されておらず，問題解決などにおいて知識を獲得する際に，出来るだけ少ない負担で，成果を得ようとする認知的経済性が得られていない [5]. また，実環境での効果検証も行われていない．そのため本研究では SAP-net を用いた転移学習手法において計算時間を低減し認知的経済性を検証することを目的とする．それに加えて，転移学習におけるエージェントや環境の対応関係を記述する手法である Inter-task mapping (ITM)[6] を用いて，実環境での活性化拡散モデルを用いた転移学習手法の効果検証も行う．

2　学習アルゴリズム

2.1　強化学習

強化学習は，エージェントが与えられた環境において試行錯誤的に行動を行うことにより，報酬が多く得られる行動を学習する手法である [1]．強化学習には Q 学習，SARSA などが存在するが，本研究では現在も研究が盛んに行われている Q 学習を採用する [7]．Q 学習における行動価値の更新式を以下に示す．

$$Q(s,a) \leftarrow Q(s,a) + \alpha\{r + \gamma \max_{b \in A} Q(s_{t+1}, b) - Q(s_t, a)\} \quad (1)$$

ここで $Q(s,a)$ は行動価値が保存されている方策，s が状態，a は行動，α ($0< \alpha \leqq 1$) が学習率，r は報酬，γ ($0< \gamma \leqq 1$) は割引率，s_{t+1} は遷移後の状態 b は遷移後の状態に対する行動，s_t は現在の状態である．α，r，γ は様々な条件下に応じて変更することが多く，全ての条件下で使用可能なパラメータは存在しなため，多種多様な条件下においてその都度パラメータを変更することが必要となる．行動価値は Q テーブルと呼ばれる Look-up table に記述され，方策として獲得される．本研究では，学習を行うにつれて最適解を獲得する行動選択関数を扱うため，十分に学習が進むまである程度のランダム行動を行い，学習が十分になったら，ほぼ学習した行動価値のみを使用するボルツマン選択を用いることとする．しかし，強化学習には学習に多くの時間が必要とされるという課題がある．

2.2　転移学習

強化学習の時間の短縮を可能とする転移学習という手法が存在する [2]．転移学習は強化学習で獲得した方策を再利用可能な状態で保存しておき，新たな環境において学習を行うエージェントに転移させることで，学習時間の削減を行う手法である．転移学習については様々な研究が行われているが，研究が多く行われているという点から，本研究では Taylor の提案した転移学習手法を採用する．転移学習における行動価値の更新式を次に示す．

$$Q_c(s_t, a_t) = Q_s(s_t, a_t) + Q_t(s_t, a_t) \quad (2)$$

$Q_s(s_t, a_t)$ は強化学習で獲得した方策，$Q_t(s_t, a_t)$ は転移先の環境で獲得する方策，$Q_c(s_t, a_t)$ は $Q_s(s_t, a_t)$ と $Q_t(s_t, a_t)$ を統合した方策である．この転移学習手法により，学習の初期段階における強化学習のようなランダム行動が減少し，学習時間の短縮を図ることが可能となる．しかし，転移学習はタスクに応じて適した知識を選択しなければ効果が得られない．そのため，強化学習で得られた方策を複数保存し，様々な環境において，効果的な方策を自律的に選択して転移学習する手法が必要であると考えられる．

3　既存研究

複数の方策を用いた転移学習を行っている既存研究として，PRQ-Learning と Strategy P そして活性化拡散モデルを用いた転移学習手法の 3 つがある．PRQ-Learning は，進行中の学習さ

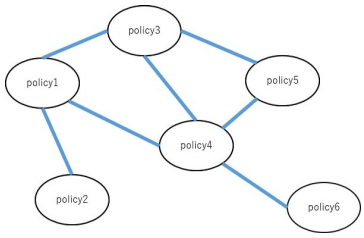

Fig.1 SAP-net

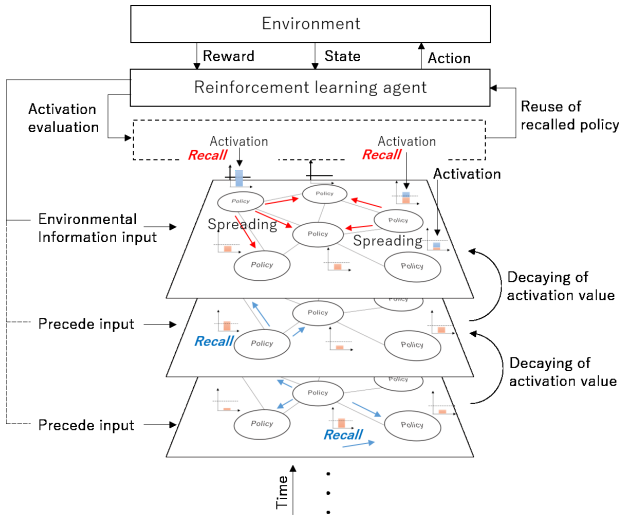

Fig.2 Transfer learning method using spreading activation model

れた方策の利用，ランダムな未踏行動の探索，過去の方策の利用の 3 つから行動選択をしながら学習を進めていく．過去の方策を利用する場合は、方策を用いて現在のタスクで行動させ，結果から方策を選択する. すべての方策で行動させなければ適した知識が選択できないため，不要な試行が必要となってしまう [8]. Strategy P は，異なる環境で同じ目的を達成するタスクである異遷移同目的タスクと、同じ環境で異なる目的を達成するタスクの同遷移異目的タスク 2 つに分け学習を進めるといったものである. 学習開始前にタスクの概念が認知できるという前提で，過去のタスクと現在のタスクを比較し，類似しているタスクの方策を転移する. この手法では、環境もタスクも変化する場合についてはどうなるかは議論されていないため、手法を使用できる範囲が限定的になってしまっている [9]．知識選択において，様々な環境やタスクに応じて適した知識を効率的に選択するために，活性化拡散モデルを用いた転移学習手法が提案されている．この手法では，認知心理学で提唱されており，ヒトの脳内にあるとされている，活性化拡散モデルを参考にしている．方策を活性化拡散モデルを参考にした図 1 のような方策ネットワーク（Spreading activation policy network: SAP-net）で定義している．方策の選択は環境入力情報とそれぞれの方策に付与された活性値を基に行われている．図 2 に Kono et al. のアプローチを示す活性値は環境入力情報を用いて活性化し，さらにそれが SAP-net を通じて関連する方策に活性値が拡散され，閾値にて想起を判断することで，環境に対して効果的な方策選択を行うというものである. しかし，実装時の方策数に応じ，計算時間が長くなるという課題がある. また，実環境における検証は行われていない [3][4].

4　提案手法

提案手法の概要としては，強化学習で得た知識に対して，それぞれの関係性を SAP-net で定義し，環境入力情報と知識ごとの活性値を元に方策を選択し転移学習をするという，活性化拡散モデルを用いた転移学習手法を用いる. 活性化拡散モデルを用いた転移学習手法の流れを以下に示す.

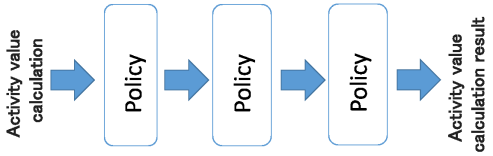

Fig.3 Existing method

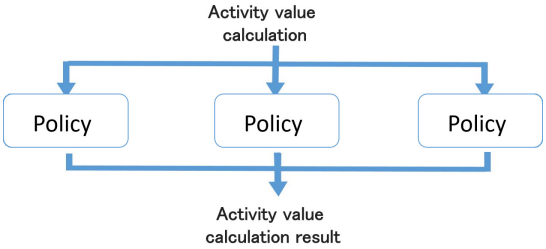

Fig.4 Proposed method

1. エージェントの周囲の環境情報を観測
2. 観測した環境情報から使用可能である方策を絞り込み，それぞれの方策に与えられている活性値情報からさらに使用可能な方策を検索する
3. 絞りこまれた方策から方策を一つのみ選択し（複数の方策候補がある場合はランダムに一つ選択），転移学習を用いて行動させる
4. 選択された方策に活性値を与え，関連した方策それぞれに対して活性値の拡散（活性化拡散）を行う
5. 転移学習の成功，失敗により，SAP-net の方策間距離の更新を行う
6. ステップ 1 へ戻る

活性化拡散モデルを用いた転移学習手法の流れにおける 2 つ目に対して，Kono et al, の手法では図 3 のように SAP-net 内に保存されている方策それぞれの個別の環境情報と活性値をのチェックをシーケンシャルに行っているため、手法実行時の計算時間が長くなる．それに対し提案手法では，既存手法の流れの 2 つ目に対して図 4 のようにシーケンシャルではなく並列化し、並列計算で動作させることに加え，並列化した方策それぞれの計算部分を常時動作させておくのではなく，必要とされたときのみ動作させることで，知識選択時間を短縮させることを提案する.

5　実験方法

5.1　シュミレーション実験

シュミレーション実験には図 5 のようなグリッドワールドを用いる．青い丸がエージェント，エージェントのいるマスが初期位置，緑のマスが目的地を示す．学習環境は，エージェントが行動可能な範囲，行動が不可能な範囲である障害物，エージェントの目的地であるゴールの 3 種類のグリッドで構成する．白のマスが行動可能範囲，黒のマスが障害物を示す．方策選択時間を測定するにあたり，10 種類の環境で獲得した 10 個の方策から方策選択に要する時間を測定する実験に加え．100 種類の環境で獲得した，100 個の方策を用いた方策選択に要する時間を測定する実験も行っていく．これにより方策数に対しての提案手法の有効性を比較する．エージェント周囲の環境情報と活性値から方策選択を行う部分の時間を比較するにあたり，方策それぞれをシーケンシャルにチェックしていく既存手法と，方策それぞれを並列化し同時にチェックする提案手法，それぞれを 50 回の知識選択を実行し，一度の方策選択にかかる時間の平均と標準偏差を算出し比較を行っていく. 実験に使用する学習パラメータを表 1 に示し，表 2 に文献 [3] にのっとり使用した SAP-net のパラメータを示

Table 1 Learning parameters

Parameter	Source-task	Target-task
Learning rate α	0.1	0.1
Discount rate γ	0.99	0.99
Boltzmann parameter ρ	0.05	0.05
Positive reward r^+	1.0	1.0
Negative reward r^-	0.0	-1.0
Number of episode	400	200

Table 2 SAP-net パラメータ

Variable	value
Default Activated value $\mathbb{A}_i$	0.0
Activation coefficient $\mathbb{A}_a$	1.0
Threshold for recall T_R	0.6
Decaying value for $\Delta\mathbb{A}_i$	-0.001
Default Weight between policies ω_{ij}	5
Adjustment value of ω_{ij} when PT	-0.5
Adjustment value of ω_{ij} when NT	2.5

す．PT は positive transfer を表し，NT は negative transfer を表している．

5.2 実機実験

実環境において，活性化拡散モデルを用いた転移学習手法の評価を行う．実験条件として，図 6 のようにグリッドワールドでの強化学習にて獲得した方策を用いる．エージェントのいるマスが初期位置，緑のマスが目的地を示しており，赤のマスが学習した経路となっている．グリッドワールドにおいては，エージェント自身の位置はグリッドワールド内のどのマスにいるかで表されている．そのため，実機実験では，約 1m × 1m のフィールドを仮想的に区切り，マス目に置き換えることで，俯瞰カメラとロボット上部に張られたマーカを用いて取得した自己位置座標を，環境情報のマス目として取得する．また．シュミレーション内での上下左右に移動する行動を，実験に用いる移動ロボットには，前進，後退，右に 90 度旋回したのちに前進し左に 90 度旋回，左に 90 度旋回したのちに前進し右に 90 度旋回の行動に対応させる．シュミレーションの場合には，エージェントは自身の周囲のマスを認識しその情報を環境入力情報としているが，実機実験では，エージェントの前後左右 50mm 以内の環境を距離センサを用いて取得し，環境入力情報とする．環境情報，行動，環境入力情報それぞれの対応関係を ITM を用いて定義する [6]．方策には環境に効果的な方策，効果的ではない方策を合計で 10 種類用意し，方策選択を行わせる．また，5 ステップ毎に環境入力情報と活性値を用いて，使用する方策を選択するものとする．実験に用いた移動ロボットを図 7 に，実験環境を図 8 に示す．移動ロボットには，計算機として Raspberry Pi，距離センサには VL53L0X を，センサ基盤には TCA9548A を用いる．ステッピングモータには，PKP24，モータドライバには DC+ Stepper Motor Hat を用いる．

6 実験結果

6.1 シュミレーション実験結果

方策 10 種類の実験結果を図 9 に．方策 100 種類の実験結果を図 10 に示す．また，それらの方策選択時間をまとめたものを表 3 に示す．方策を 10 種類用いた実験において，既存手法では方策選択に必要な時間が平均 0.253ms とり提案手法では 0.091ms となっていることから約 65 %の時間を削減出来ていることが分かる．次に，方策 100 種類において，既存手法では方策選択に必要な時間が平均 1.738ms となり提案手法では 0.438ms となり約 75 %の時間を削減出来ていることが分かる．これらのことから

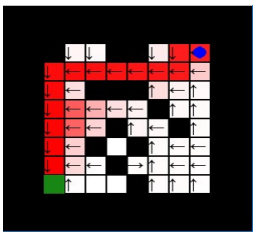

Fig.5 Grid world example

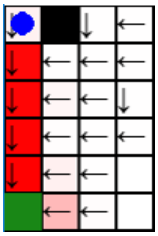

Fig.6 Example of learned policy

今回の提案手法は方策の種類が増えるほど効果を発揮すると考えられる．

6.2 実機実験結果

実機での強化学習と活性化拡散モデルを用いた転移学習手法との比較をする学習曲線を図 11 に示す．学習においてタスクの達成回数を Episode, 行動回数を Step と呼ぶ．活性化拡散モデルを用いた転移学習手法っでの実験は 5 Episode までとなっている．また，活性化拡散モデルを用いた転移学習手法にて得られた活性値を図 12 に示す．図 11 から強化学習にくらべ初期の段階から非常に少ない Step 数でタスクを達成していることが分かる．今回は環境に対し有効な方策を選択し，行動することで最短経路でゴールするように方策を用意したため，初期から環境に適した方策を選択し最短経路でゴールしていることが確認できる．また，図 12 から活性値が上昇している方策が環境に効果的な方策だと考えられる．これらの事より活性化拡散モデルを用いた転移学習手法は実環境においても有効だと考えられる．

7 結言

本論文では，活性化拡散モデルを用いた転移学習手法に対して，方策の種類が増えるほどに方策選択時間が増加してしまうことに加え，実環境において検証されていないことを指摘した．それらに対して，方策選択時間を短縮するためにシーケンシャルにチェックしていく方法から並列的に動作させること，実環境においての検証を行うことを提案し，実験を行った．方策選択時間に関しては，提案手法を用いることで，方策 10 種類の場合には約 65 %の削減，方策 100 種類の場合には約 75 %の削減を確認することができ，提案手法が方策の種類が増えるほど効果を発揮する事を確認できた．実環境においての検証ではグリッドワールドで方策を獲得し，実環境にて転移学習を行った．その結果から活性化拡散モデルを用いた転移学習手法は実環境においても有効だと確認できた．シュミレーション実験では方策を 10 種類と１００種類の 2 つのパターンで検証を行ったが，今後はさらに方策の種類を増やし提案手法の有用性の確認を行っていく．また，実環境においても方策の種類も少なく限定的なため，環境を変化させたパターンや方策の種類を増やした実験も行っていく．

謝辞

本研究は JSPS 科研費 JP19K12173 の助成を受けたものである．

参考文献

[1] R. S. Sutton, A. Gbarto（三上貞芳，皆川雅章訳） :強化学習, 森北出版, 2000.

[2] Taylor, M. E. and Stone, P: “Transfer learning for reinforcement learning,” J. Machine Learning Research, vol.10, no.10, pp.1633–1685, 2009.

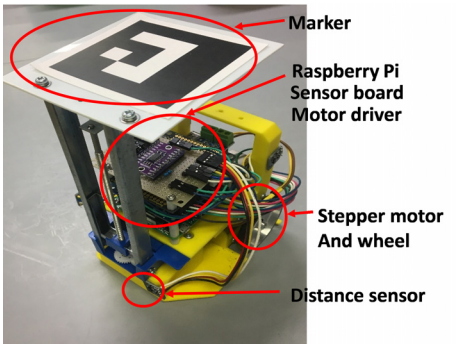

Fig.7 Mobile robot

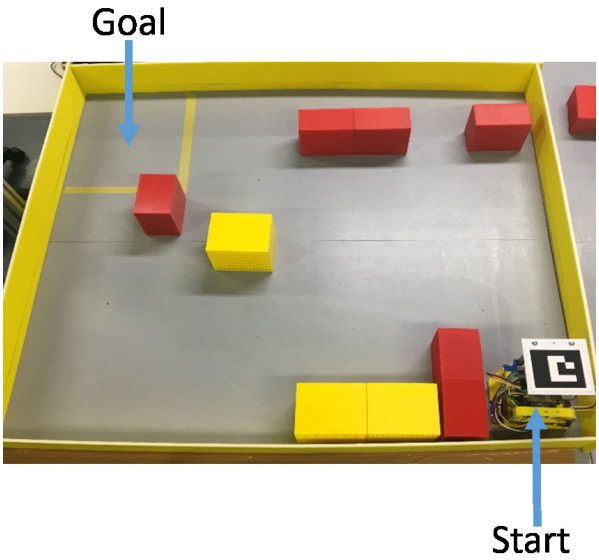

Fig.8 Experiment environment

[3] Hitoshi Kono, Ren Katayama, Yusaku Takakuwa, Wen Wen, Tsuyoshi Suzuki: "Effenctive Activation and Speading Sequence for Transfer Learning in Reinforcement Lerning with Spreading Activation Policy Network," International Journal of Advanced Computer Science and Applications, vol.10, no.12, 2019.

[4] 高桑優作, 鈴木剛, 河野仁, 温文: "活性化拡散モデルに基づく強化学習エージェントの方策選択手法," ロボティクス・メカトロニクス講演会講演, 2P2-E04, 2017.

[5] 野津亮, 山本優, 本多克宏, 市橋秀友: "認知的経済性に基づいた社会シミュレーションモデルにおけるコミュニケーション形態の影響," 知能と情報, vol.22, no.2, pp.154–164, 2010.

[6] M. E. Taylor, P. Stone, and Y. Liu: "Transfer learning via inter-task mappings for temporal difference learning," J. Machine Learning Research, vol.8, no.1, pp.2125–2167, 2007.

[7] C. J. C. H. Watkins and P. Dayan: "Q-learning," Machine Learning 8, pp.279–292, 1992.

[8] Fernando Fernandez, Manuela Veloso : "Probabilistic Policy Reuse, Progress in Arti-ficial Intelligence, vol.2, no.2, pp.13–27, 2013.

[9] Toshiaki Takano, Haruhiko Takase, Hiroharu Kawanaka Hidehiko Kita, Terumine Hayashi and Shinji Tsuruoka : Transfer learning based on forbidden rule set in actor-critic method, International Journal of Innovative Computing, Information and Control, vol.7, no.5(B), pp.2907–2917 2000.

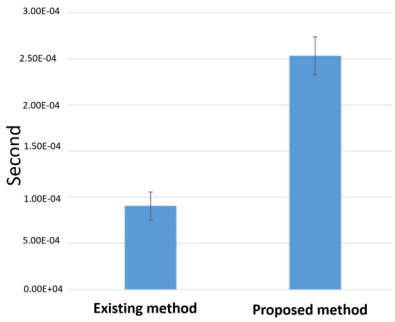

Fig.9 Policy selection time with 10 types(n=50)

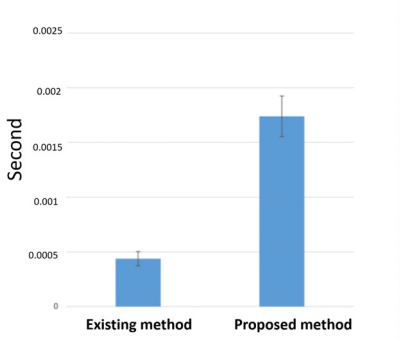

Fig.10 Policy selection time with 100 types(n=50)

Table 3 Policy selection time(n=50)

	10 Policies	100 Policies
Existing method[3]	0.253ms	1.738ms
Proposed method	0.091ms	0.438ms

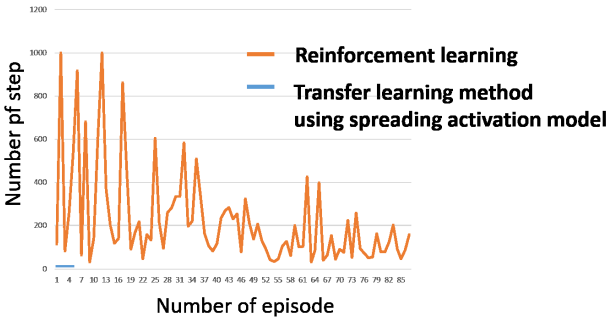

Fig.11 Learning curves result in actual mobile robot

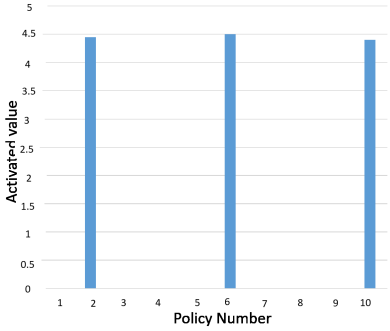

Fig.12 Activated value of policies when simulation is terminated.