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

○ 高桑 優作（東京電機大） 河野 仁（東京工芸大学） 温 文（東大）
神村 明哉（産総研） 富田 康治（産総研） 鈴木 剛（東京電機大）

Policy Selection Method based on Spreading Activation Model for Reinforcement Learning Agent

○Yusaku TAKAKUWA (Tokyo Denki University), Hitoshi KONO (Tokyo Polytechnic University), Wen WEN (University of Tokyo), Akiya KAMIMURA (AIST), Kohji TOMITA(AIST), and Tsuyoshi SUZUKI(Tokyo Denki University)

Abstract: This paper proposes a policy selection method of a reinforcement learning agent for suitable learning in unknown or dynamic environments based on a spreading activation model in the cognitive psychology. The reinforcement learning agent saves policies learned in various environments and the agent learns flexibly by partially using suitable policy according to the environment.

1. 緒言

近年，ロボットを未知の環境及び動的環境のような実環境への適応を可能にすることを目的とした研究が盛んに行われている[1]．全ての状態を想定した実環境に対してロボットを適応的に行動させるために，人が制御則を与えることは困難であることから，ロボットに自律的に行動を学習させる強化学習，転移学習が用いられている．

強化学習とは，エージェント(以下，学習可能なロボットをエージェントと呼称)に試行錯誤を行わせることで，ある環境における最適な行動を学習させる手法である[2]．また，転移学習は，獲得済み知識を再利用することで新たな環境への適応能力向上や学習時間短縮を図る手法である[3]．

エージェントを未知の環境や動的な環境に適応させるためには，単一の知識ではなく複数の知識を利用することが必要と考えられる．既存の研究においても複数の学習知識を保存し選択して使用した場合の検討がなされているが[4,5]，エージェントが観測した情報に基づく選択手法ではなく，動的環境や未知の環境における学習は考慮されていない．そのため，複数の知識をエージェントが観測した情報に基づいて選択する手法の検討が必要である．

本研究では，エージェントが獲得・保存した複数の識(以下，強化学習知識を方策と呼称)の再利用における方策選択手法の確立を目指す．人間は学習などで得た知識・記憶の選択により判断や行動を行っていることが認知心理学的知見[6]として得られていることから，本稿では認知心理学の知見である活性化拡散モデルを用いた方策間関係の記述及び方策選択の手法を提案し，計算機シミュレーションにより検証する．

2. 活性化拡散モデル

活性化拡散モデルとは，人間が獲得した概念同士が脳内でネットワーク構造として保存されていることを前提とし，ある概念が想起されることで，関連する概念も活性化され，概念利用が促進されるモデルである[7]．

活性化拡散モデルは，関連性の強さに応じて概念間の距離を変動させて配置する意味的距離を定義する．活性化拡散モデルの例を図1 に示す．図1の棒グラフは，概念毎に与えられた活性値を表し，この値が閾値を超えることで，概念が想起される．概念の活性化は，構築されたネットワークを通じて行われる．

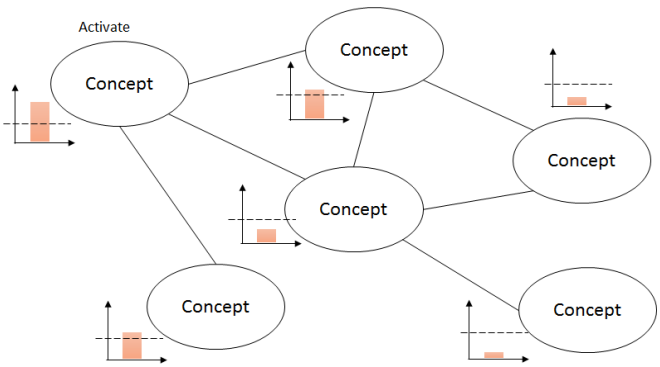

Fig. 1 Spreading activation model (activation of the concept is done through the constructed network)

3. 提案手法

3.1. 前提条件

本研究では強化学習するエージェントが獲得した方策を選択するメカニズムを提案する．本節では，提案手法を述べるにあたり，関係用語や前提を述べる．

強化学習には，Q 学習を用いる．状態s における行動a の価値である行動価値$Q(s_t,a)$ (以下，Q 値と呼称)の更新式を次に示す．tは状態の遷移前，$t+1$は状態の遷移後を表す．

$$Q(s_t,a) \leftarrow (1-\alpha)\,Q(s_t,a)+\alpha[r_{t+1}+\gamma \max Q(s_{t+1},a)] \qquad (1)$$

ここで，α を学習率，γ を割引率，r を報酬とする．学習したQ 値及び学習した環境における環境情報（障害物などの情報）は，Q-table と呼ばれるLook-up tableを用いて記述しておき，予め再利用可能な状態で保存する．方策再利用に関する転移学習の更新式を次に示す[3]．

$$Q_c(s_t,a) \leftarrow \tau\, Q_s(s_t,a)+Q_t(s_t,a) \qquad (2)$$

転移学習では，学習予定のタスク(以下，Target-task と呼称)に対して，予め学習したタスク(以下，Source-task と呼称)で獲得した方策を転移する．$Q_s(s_t,a)$ は Source-task 方策，$Q_t(s_t,a)$ は Target-task で学習中の方策，$Q_c(s_t,a)$ は統合した方策を示している．τ は利用方策の Q 値を調整するパラメータ(以下，転移率と呼称)である．

提案手法では，活性化拡散モデルの意味的距離を参考に，予め全ての方策間を方策間距離 d_{nm} という有向グラフで結合する．使用中方策の識別番号は n，距離の結合先の方策の識別番号は m で示している．また，同じ方策の連続使用を考慮し，d_{11} や d_{22} のような同じ方策を結合する方策間距離も付与する．方策間を方策間距離で結合済みの方策間関係（Spreading Activation Policy Network : SAP-Net）の例を図 2 に示す．πは，学習済みの方策を表している．

3.2. 提案手法の流れ

エージェントは，タスクの達成のため自身の周囲の環境情報を観測し，SAP-Netと観測した環境情報を基に，方策を選択して転移する．エージェントは，方策の想起，行動，活性化，方策間距離のフィルタリングという機能を反復しながら学習する．この転移学習の目的は，予め学習した方策を手掛かりにTarget-taskにおける方策を獲得することである．

3.3. 方策間距離を基にした方策の想起

エージェントが観測した環境情報及びSAP-Netを参照して確率的に方策間距離を選択することを本研究において想起と呼ぶ．選択した方策間距離の結合先の方策を転移する方策として選択する．

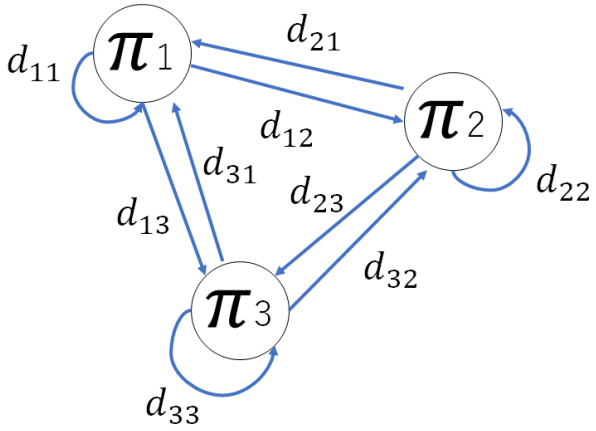

Fig. 2 SAP-Net configured multiple polices linked distance of policies

方策の選択に用いた方策間距離は，次節で述べる活性化処理を施す．使用する方策は，方策の選択毎に切り替える．

3.4. 方策間距離の活性化

活性化処理では，想起により方策が適切に選択され転移された場合（以下，正の転移と呼称）に方策間距離を縮め，適切でない転移の場合（以下，負の転移と呼称）に，方策間距離を伸ばす．この処理により，正の転移がなされる方策間の関連性を記述する．方策間距離 d_{nm} の更新式を次に示す．

$$d_{nm}=\begin{cases} d_{nm}-e_{activate} & (\text{正の転移}) \\ d_{nm}+e_{attenuate} & (\text{負の転移}) \end{cases} \qquad (3)$$

式の $e_{activate}$(以下, 活性化係数と呼称)と $e_{attenuate}$(以下, 減衰係数と呼称)は任意定数とし，方策間距離を変動する．使用中方策を π_n，選択する方策 π_m，使用中方策から伸びる方策間距離の本数を i とした時，方策間距離 d_{nm} を通じて使用中方策 π_n から選択する方策 π_m を選択する確率 $P(\pi_m \mid \pi_n)$ を求める式を次に示す

$$P(\pi_m \mid \pi_n)=\frac{(d_{nm})^{-1}}{\sum_i (d_{nm})^{-1}} \qquad (4)$$

3.5. 方策間距離のフィルタリング

方策間距離の大きさに閾値を設け，閾値を超えた負の転移を促進する方策間距離は，想起で参照されないように設定する．この処理により，正の転移を促進する方策を判定し転移する．

4. メタデータの利用

本研究では，エージェントが Source-task 学習中に副次的に獲得した情報(以下，メタデータと呼称)を，方策選択の基準及び，活性化処理における正の転移の判定に利用する．メタデータを用いた転移学習の更新式を次に示す．

$$Q_c(s_{target}, a) \leftarrow \tau Q_s(s_{meta}, a) + Q_t(s_{target}, a) \qquad (5)$$

s_{target}は Target-task におけるエージェントの状態，s_{meta}は Source-task におけるメタデータを観測した状態を示す． 本研究で用いるメタデータは，Source-task に依存するため，学習させるタスクに応じて設定する．

5. 計算機実験

本実験では，提案手法による転移学習と強化学習との比較を行う． 比較する点は学習に必要となった行動数の合計である総 Step 数，学習初期における行動回数の減少(以下，ジャンプスタートと呼称)，学習におけるエージェントの最適行動回数への収束という 3 つの指標で評価を行う．

5.1. 実験設定

エージェントが学習を行う動的環境のタスクとして追跡問題を用いる．追跡問題は，ハンター側のエージェントが獲物を追跡し，捕らえるとハンターに報酬が与えられるタスクである．エージェントの行動機構は，上下左右への移動に停止を加えた 5 つの行動を可能にするものとする．獲物は，上下左右 2 マスをセンシングし，ハンターを発見した場合にハンターから離れるように行動する．

学習環境はグリッドワールドに設定する．追跡問題をグリッドワールドで表現した図を図 3 に示す．エージェントの行動可能範囲は，6×6 マスとする．

タスクの達成回数を Episode 数，各 Episode に対するエージェントの行動回数を Step 数と呼称する．学習は，1-999episode まで行う．

5.2. メタデータの設定

エージェントが Source-task 学習時に，獲得可能なメタデータは，エージェント自身と獲物の初期座標及び学習進行後(Episode949-999)における Step 数の平均値とする．エージェントは，行動毎に自身と獲物との相対座標を観測できるものとし，方策の想起では環境情報として取得した相対座標とメタデータとして取得した Source-task の相対座標を照合して，方策の選択候補を検索する．活性化処理では，方策を転移後の行動回数を観測し，Source-task 学習時の Step 数以内で獲物を捕らえた場合を正の転移とする．

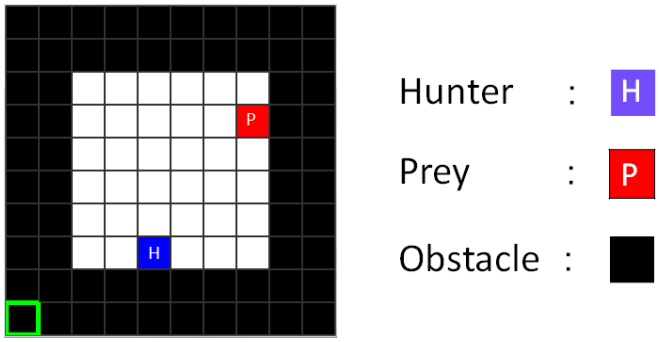

Fig.3 Tracking problem

5.3. 交差検証

本実験では，ランダムにハンターの学習初期座標，獲物の初期座標を設定した 100 パターンの環境を予め Source-task として学習し，方策を再利用可能な状態で保存する．保存した方策は，3 章で述べた SAP-Net 構築に用いる．方策はπ_1からπ_{50}，π_{51}からπ_{100}までの 2 集合に分割し，各集合内の方策を用いて転移学習を行う．転移学習を行う環境は，π_1からπ_{50}を用いる場合は環境 51 から環境 100 を，π_{51}からπ_{100}を用いる場合は環境 1 から環境 50 を学習する．この交差検証によって，複数の学習サンプルを基に統計的に学習効率を検証する．方策と学習環境の交差検証を示した図を図 4 に示す．100 の環境の中で，総 Step 数，ジャンプスタート，収束の指標から，強化学習と比較して転移学習が効率的であった環境を数える．効率的だった環境の数を用いて，全体の環境の中で提案手法による転移学習が効率的だった環境を割合で算出する．

5.4. 評価指標の設定

評価指標となる総 Step 数，ジャンプスタート，収束についての詳細を以下に示す．

・学習に必要となった行動数の合計を総 Step 数とする．

・学習初期(Episode1-50)における Step 数の平均を比較して，Step 数が減少していればジャンプスタートを観測したものとする．

・学習進行後(Episode949-999)における Step 数の平均を比較して，Step 数が減少していれば最適 Step 数への収束が進行しているものとする．

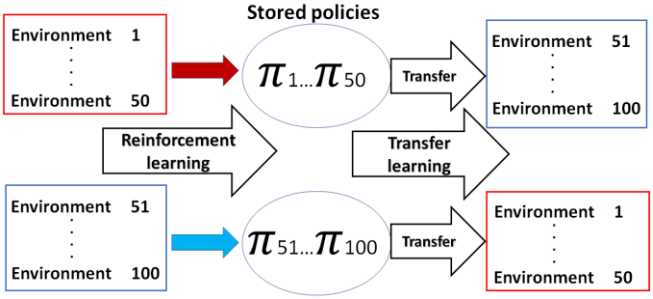

Fig. 4 Cross validation of transfer learning

転移学習に用いたパラメータの設定を表 1 に示す.

5.5. 実験結果

表 2 に実験結果を示す．表 2 から，強化学習と転移学習との比較において，どの評価指標においても転移学習による学習効率が上回った環境は，全体の半数以上確認された．図 5 に評価指標に基づき正の転移が多く観測された環境における学習曲線(Episode900-999)の例を示し，図 6 に負の転移が多く観測された環境における学習曲線の例(Episode900-999)を示す．図 5 から，強化学習と比較して，転移学習による Step 数の減少が確認できる．一方，図 6 では，強化学習よりも転移学習の Step 数が伸びていることが確認できる．正の転移が観測されれば，Step 数は減少するが，負の転移が観測されれば，Step 数は増加する．この結果の原因は，予め学習した方策が，図 6 の学習環境には不適であったためであると推測される．また，表 2 の結果においても，エージェントが観測した環境情報に合う方策が Source-task で学習した方策の中に存在しないことから，全ての環境で効率的な学習ができなかったと推測される．本研究は，予め学習した方策を再利用するため，転移学習は Source-task で学習した方策に依存する．提案手法における転移学習を用いて，正の転移を促進するためには，Source-task で学習する方策を増加し，Target-task に適する方策を確保することが重要であると考えられる．

6. 結言

本稿では，複数方策を方策間の関係と転移先の環境に応じて選択する手法を提案した．実験では，交差検証を用いて提案手法による転移学習と強化学習とを比較し，方策及び学習環境の偏りに関係なく複数方策を用いた転移学習が可能であることを示した．今後は，実機ロボットによる学習実験及び，エージェントが獲得した方策数と提案手法を用いた際の学習効率の関係性について検討していく．

謝辞

本研究の一部は科研費(16K12493)の助成を受けて行われた．ここに謝意を表する．

引用文献

1) 赤井直紀，山内　健司，井上一道，宇内　隆太郎，山本　条太郎，尾崎　功一：“つくばチャレンジ 2013 の課題を題材とした実環境におけるタスク遂行ロボットの開発”，計測自動制御学会論文誌，Vol.51，No.1，24/31(2015).

2) 三上　貞芳，皆川　雅章：”強化学習”，3-5，森北出版株式会社(2000).

3) M.E.Taylor ： “Transfer in Reinforcement Learning Domains” , 122, Springer(2009) .

4) F.Fernandez , M.Veloso ： “ Probabilistic Policy Reuse in a Reinforcement Learning Agent ” , *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems AAMAS'06*, May 8-12, Hakodate, Hokkaido. Japan，2006 .

5) 吉田　慎二，長谷川　修：“強化学習における政策再利用転移学習”，情報処理学会第 73 回全国大会，2011.

6) 安藤　清志，石口　彰，高橋　晃，浜村　良久，藤井　輝男，八木　保樹，山田　一之，渡邊正孝，重野　純，“キーワードコレクション　心理学　改訂版”，210-231．新曜社，2012.

7) A. M. Collins，E. F. Loftus ：“A Spreading-Activation Theory of Semantic Processing”, *Psychological Review*，Vol. 82，No. 6，407-428(1975) .

Table. 1 Parameters setting

Parameter	Symbol	Value
Learning rate	α	0.2
Discount rate	γ	0.9
Reward	r	1.0
Transfer rate	τ	0.3
Activate coefficient	$e_{activate}$	1.0
Attenuate coefficient	$e_{attenuate}$	1.0

Table. 2 Experimental result

	Proportion of environments which efficient transfer learning
Total number of steps	80 %
Jump start	77 %
Convergence of steps	61 %

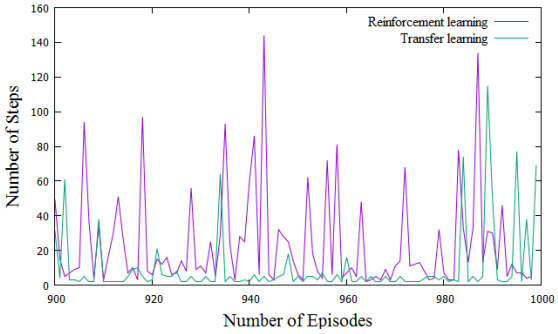

Fig. 5 An example of the learning curve in which a positive transition is exhibited (Episode900-999)

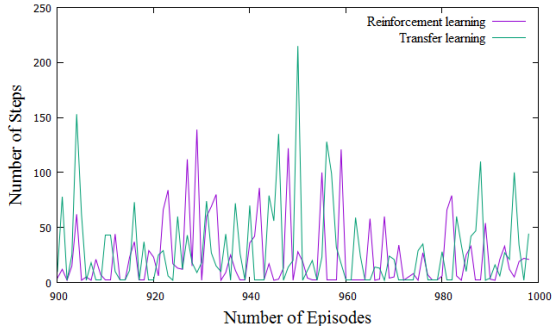

Fig. 6 An example of the learning curve in which a negative transition is exhibited (Episode900-999)