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

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

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

Yusaku TAKAKUWA, Tokyo Denki University, y.takakuwa@nrl.c.dendai.ac.jp
Hitoshi KONO, The University of Tokyo
Wen WEN, The University of Tokyo
Akiya KAMIMURA, National Institute of Advanced Industrial Science and Technology (AIST)
Kohji TOMITA, National Institute of Advanced Industrial Science and Technology (AIST)
Tsuyoshi SUZUKI, Tokyo Denki University

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. In the proposed method, a directed graph is created between policies, and the network is constructed by means of a policy by combining them between policies. The agent updates the network according to the environment while repeating processes of recall, activation, filtering, and learns based on the network. Agent uses this network in transfer learning. Simulation results show that reinforcement learning agent achieves task by selecting the optimal one from multiple policies by the proposed method and from the comparison of transfer learning with the proposed method and the learning efficiency of ordinary reinforcement learning, the usefulness of the proposed method.

Key Words: Reinforcement learning, Mobile robot, Spreading activation

1. 緒言

近年，学習能力，認識・理解能力などの人や動物が持つ知的能力を有するロボットの実用化が期待されている．ロボットを知能化し，未知の環境及び動的な環境へ適応を可能にする研究が盛んに行われている[1].

未知の環境や動的な環境へ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，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．しかし，既存の研究では複数の学習知識を保存した場合の使用についての検討が殆どない．そのため，複数の知識を学習毎に個別に保存し，知識を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・保存した複数の知識(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，概念の利用が促進されるモデルである[4].

活性化拡散モデルには，関連性の強さに応じて関連している概念間の距離を変動させて配置する意味的距離が存在する．概念の活性化は，関連性によって構築されたネットワークを通じて行われ，各概念間に意味的な関連性の表現が存在する．活性化拡散モデルの例を図 1 に示す[4].

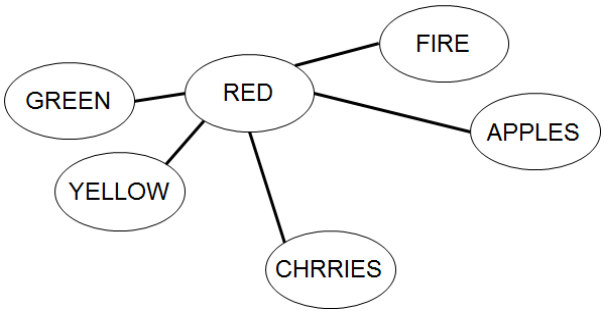

Fig. 1 Spreading activation model (where a shorter line represents greater relationship)

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．時刻 t での状態 s における行動 a の価値である行動価値 $Q(s_t, a_t)$(以下，Q 値と呼称)の更新式を次に示す.

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

ここで，s_t, a_t は現在の状態及び行動，s_{t+1} は行動 a_t を行った後の状態，α を学習率，γ を割引率，r_{t+1} を報酬とする．

初期位置から目的地までの最短の経路を学習する最短経路問題を Q 学習のタスクとして設定する．

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．提案手法で

は，この Q-table を方策として利用する．方策の再利用に関する転移学習の更新式を次に示す．

$$Q_c(s_t,a_t) \leftarrow Q_s(s_t,a_t)+Q_t(s_t,a_t) \tag{2}$$

$Q_s(s_t,a_t)$ は再利用方策，$Q_t(s_t,a_t)$ は Target-task で学習中の方策，$Q_c(s_t,a_t)$ は統合した方策を示している．

提案手法では，あらかじめ方策同士を方策間距離 d_{nm} という有向グラフで結合する．n は使用中方策の識別番号を，m は距離の結合先の方策の識別番号を示す．また，同じ方策の連続使用を考慮し，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（Policy network : Pnet）の例を図 2 に示す．K は，学習済みの方策を表している．

3.2 提案手法の流れ

エージェントは自身の周辺の環境情報を観測し，Pnet と観測した情報を基に，方策を選択して使用する．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順に反復して行いながらタスクの達成を図る．

3.2.1 方策間距離を基にした方策想起

方策の想起では，各方策間距離の持つ想起確率を参照して確率的に方策間距離の選択を行う．選択した方策間距離の先に結合する方策を次の使用方策として選択する．

選択して使用した方策間距離には，次節でのベル活性化処理を施す．使用する方策は，方策の想起毎に切り替える．

3.2.2 方策間距離の活性化

活性化処理では，想起が適切に行われた場合に方策間距離を縮め，適切に行われなければ方策間距離を伸ばす．方策間距離 d_{nm} の更新式を次に示す．

$$d_{nm} = \begin{cases} d_{nm} - e_1 & (\text{想起が適切}) \\ d_{nm} + e_2 & (\text{想起が不適切}) \end{cases} \tag{3}$$

この式の e_1，e_2 は任意定数とし，方策間距離を変動する．使用中方策を K_n，選択する方策を K_m，使用中の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 K_n から選択する方策 K_m を想起する確率 $P(K_m \mid K_n)$ を求める式を次に示す．

$$P(K_m \mid K_n) = \frac{(d_{nm})^{-1}}{\sum_i (d_{ni})^{-1}} \tag{4}$$

3.2.3 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，閾値を超えた方策間距離はフィルタリングをして，想起に使用されないように設定する．この処理により，適切な想起が見込まれる方策間距離を参照し続けることで学習する．

4. 計算機実験

本実験では，提案手法による転移学習の正常動作と，提案手法による転移学習と強化学習との比較を行う．

提案手法による転移学習は，環境情報に合う方策を選択し転移することで学習する．この実験では，Target-task における Q 値の獲得を目的とする．

二つの学習手法の比較実験では，タスクの達成回数である Episode 数と試行回数である Step 数を基にグラフ化し，学習効率を比較する．

4.1 計算機実験の設定

エージェントは，上下左右への移動に停止を加えた 5 つの行動が可能であるものとする．また，方策は，6 つの学習環境 (Source-task) で学習したものを用いる．学習予定の環境 (Target-task) を図 3 に，6 つの方策を学習した学習環境 (Source-task) を図 4 に示す．図 3，図 4 の S は初期位置，G は目的地，白いマスは行動可能マス，黒いマスは障害物を示す．図 3 の矢印は，学習時の方策を示す．図 3 の A はエージェントを示す．本実験に使用する 6 つの方策の中には，使用するとゴールへの接近行動を妨げる非効率的な方策も用意する．

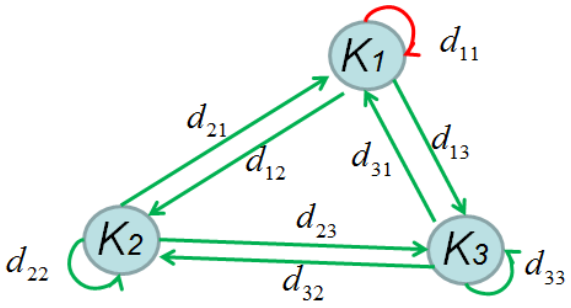

Fig. 2 Pnet configured multiple policies linked distance of policy

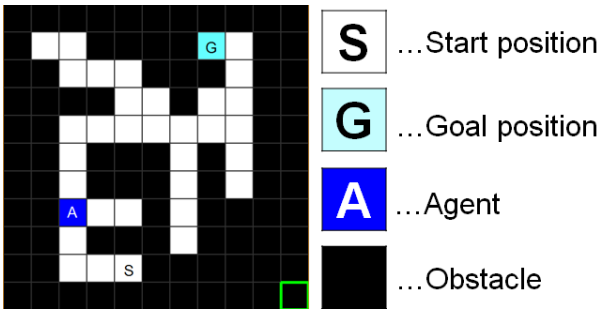

Fig. 3 Target task

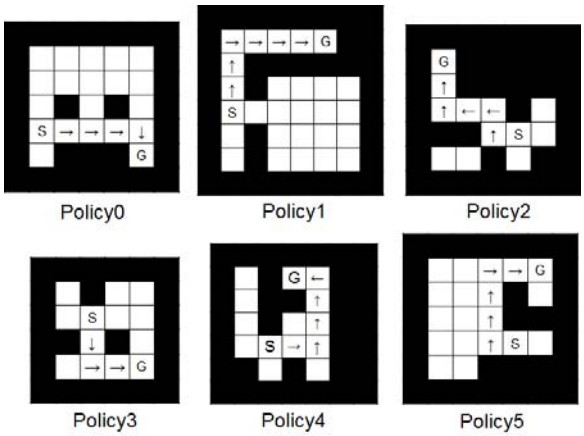

Fig.4 Six policies that agent obtained in reinforcement learning

Table 1 Parameters setting

Parameter	Symbol	Value
Learning rate	α	0.1
Discount rate	γ	0.9
Reward	r	1.0
Initial value of policy distance	d_{nm}	3
Initial recall probability	$P(K_m \mid K_n)$	1/6

提案手法における適切な想起の判定基準は，想起した方策による行動が障害物に妨げられないこととする．前提条件で説明した手法により方策間距離を用いて，各方策間を結合して得られた Pnet を図 5 に示す．図 5 に記述された方策間距離の大きさや想起確率の初期値は基準となる一定値に統一し，学習が進むことにより知識が想起されると，方策間距離と想起確率が活性化処理により調整される．強化学習に使用したパラメータ，方策間距離の大きさの初期値や想起確率の初期値を表 1 に示す．

シミュレータでは, Target task における学習で獲得した Q 値を，Q 値が存在する座標毎に赤色で描画することによって表現する．学習によって Q 値が高くなるにつれて濃く描画することで，Q 値の伝搬の様子を表現する．

4.2 計算機実験の流れ

エージェントは，自身の周囲 1 マスの環境情報を行動ごとに常に取得し，学習済みの各方策に存在する初期位置の環境情報と照合する．照合した方策の中から，提案手法を用いて使用する方策を選択する．使用する方策から $Q_S(s_t, a_t)$ を取り出し，式(2)を用いて統合方策 $Q_C(s_t, a_t)$ を生成する．エージェントは，統合方策 $Q_C(s_t, a_t)$ と行動選択関数を基に行動する．使用した方策は，一時的に記憶しておき次の想起に利用する．想起に伴う活性化により使用した方策間距離の大きさと想起確率を調整する．フィルタリング処理は，エージェントの行動毎に行う．

4.3 実験結果

提案手法を用いた転移学習の正常動作を確認する実験において，学習予定の環境の Q 値伝搬の様子を図 6 に示す．図 6 から Target-task における Q 値が最短経路に伝搬していることが確認できる．このことから，提案手法による転移学習が正常に行われていることを確認した．この学習によって変動した Pnet を図 7 に示す．図 5 と図 7 から方策間距離の変動を用いてフィルタリング処理が行われ，参照される方策間距離が減少したことを確認した．

このことから，想起が適切に行われ，使用すべき方策の選択が行われやすくなったことが確認できた．

提案手法による転移学習と強化学習の比較実験における学習曲線を図 5 に示す. 図 5 から転移学習をした方が Episode 達成に必要な Step 数が減少していることが確認できた．これより，効率的な学習を確認し，提案手法の有用性が示された．

5. 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，提案手法を用いた転移学習にて環境に応じて方策選択を行い，提案手法の有用性を確認した．今後は，最短経路問題以外のタスクに対して適用可能な想起の判定基準など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．

参考文献

[1] Y.C.Choi, H.S.Ahn, ”A survey on multi-robot reinforcement learning : Coordination problems”, *Proceedings of 2010 IEEE / ASME International Conference on Mechatronic and Embedded Systems and Applications*, pp.81-86, 2010.

[2] 三上　貞芳, 皆川　雅章, ”強化学習”, ISBN4-627-82661-3, 2000.

[3] M.E.Taylor, “Transfer in Reinforcement Learning Domains” , Springer, 2009.

[4] A. M. Collins, E. F. Loftus, “A Spreading-Activation Theory of Semantic Processing”, *Psychological Review*, Vol. 82, No. 6, pp.407-428, 1975.

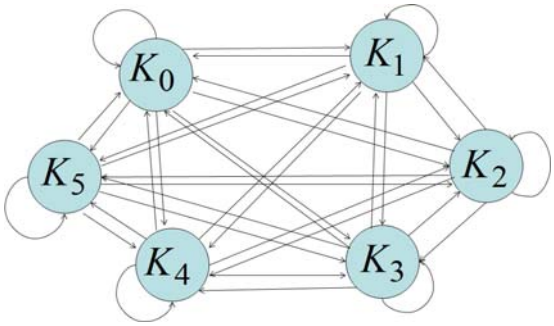

Fig. 5 Simplified overview of the default Pnet (where all distances between policies are set to same value)

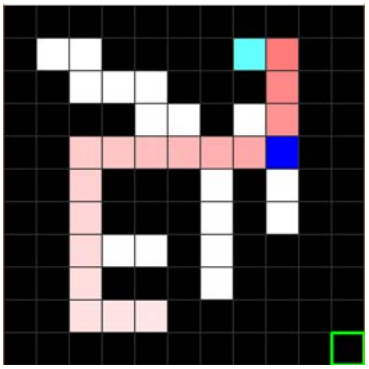

Fig. 6 Propagation of Q-value (red grids represent parts of Q-value)

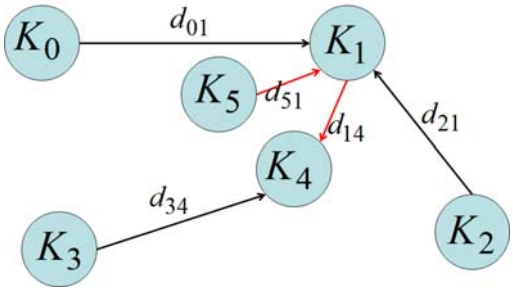

Fig. 7 State of the Pnet after the experiment (a shorter distance between policies represents strong relatedness in the similar learning environments)

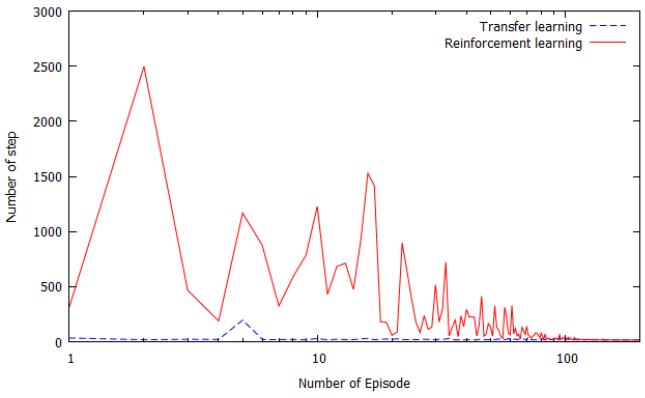

Fig. 8 Comparison in learning curves