転移強化学習における活性化拡散選択手法および拡散シーケンス

IV.提案する手法

このセクションでは、提案する手法について説明します。メソッドの関連パラメータは表Iに定義されています。後のサブセクションでは、提案されたメソッドで使用される必要な関数について説明します。予備的な関数の説明から始めます。エージェントには初期状態の方策と方策ネットワーク構造があり、すべての方策にはさまざまな活性値があると想定しています。

1) エージェントは、センサーを介して環境情報を観測します。
2) 観測された環境情報から特徴を抽出します。
3) 抽出された特徴と、以前学習した環境情報を使用してラベル付けされた方策のラベルとの一致に基づいて、対応する方策が活性化します。
4) 活性値は、活性化された方策から近くの方策に広がります。
5) 候補方策は、しきい値を使用して収集されます。
6) 確率関数に基づく想起方策の選択します。
7) 転移方策（この部分は転移学習です。）。
8) 行動と学習の選択。（この部分は強化学習です）。
9) 方策の再利用の有効性の評価します。
10) 方策ネットワーク構造の重みの調整をします。
11) 1）のプロセスに戻ります。

上記から、強化学習を備えた提案手法の簡略化されたシステムアーキテクチャを図2に示します。

A.活性化拡散方策ネットワーク

転移学習を通した強化学習法による方策選択手法を選択するために、拡散活性化モードが採用されています。活性化拡散方策ネットワーク（SAP-net）は、強化学習における転移学習の方策選択方法です。これは、Takakuwaらによって提案されました。 [27]。Takakuwaらをもとにした効果的な機能を提案します。

初期化では、無向グラフGはG = {V、E}によって定義されるため、取得されたポリシーπiは方策の集合に含まれます。ここで、Eはポリシー間のエッジのセットを示します。SAP-netは、Gによって与えられる隣接行列Aと重みのセットWとして定義されます。なので、A =（G、W）となります。

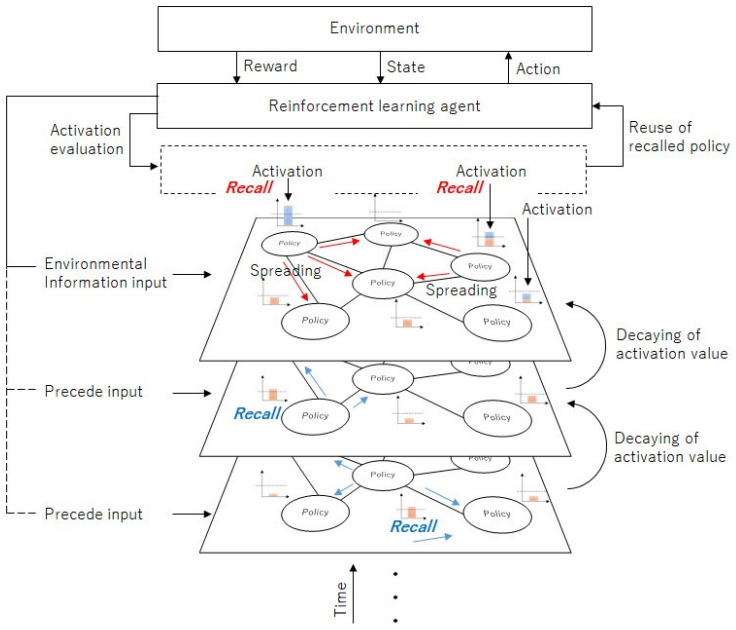

図2.提案された方法の簡略化されたシステムアーキテクチャ。方策選択のための活性化拡散手順を伴う標準的な強化学習の概念の概要を表します。

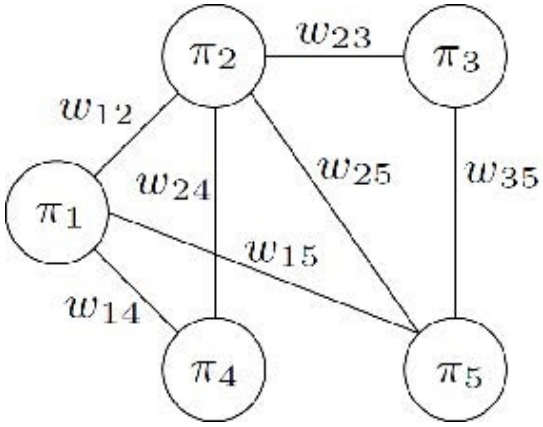

図3. SAP-netの例。デフォルトでは、すべてのノードがフルメッシュで接続されています。すべてのパスwijには重み値が割り当てられ、すべてのπkにはアクティベーション値Akがあります。

$$\boldsymbol{A} = \begin{array}{c} \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{array} \begin{array}{c} \begin{array}{ccccc} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \end{array} \\ \begin{pmatrix} 0 & w_{12} & 0 & w_{14} & w_{15} \\ w_{12} & 0 & w_{23} & w_{24} & w_{25} \\ 0 & w_{23} & 0 & 0 & w_{35} \\ w_{14} & w_{24} & 0 & 0 & 0 \\ w_{15} & w_{25} & w_{35} & 0 & 0 \end{pmatrix} \end{array} \tag{8}$$

上記の式は、wij∈Wおよびwij≥1の場合の図（3）に関連する隣接行列の例です。wijの値が大きいと、パスが長くなることに注意してください。初期状態Gはフルメッシュネットワーク

であり、グラフの動作は重みwijに依存します。非常に大きなwijは、パスが漸近的に切断されていることを示します。

すべての方策は活性化された値Aを持っています。したがって、グラフのノードvi∈Vは（πi、Ai）として構成されます。デフォルトとしてAi = 0となります。

B.活性化機能

対応する方策とペアになった機能と一致して、活性化された値Aiは、観測された環境情報の特徴間の比較関数C（・、・）によって更新されます

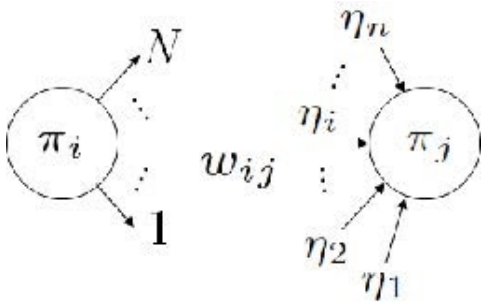

図4.πiからπjへの活性化された値の拡散の簡略図。この状況では、πiが活性化され、その値を出力します。 πjは、wijによって削減される活性化された値ηiを受け取ります。

sと方策の特徴は方策取得時の環境です。活性化機能は、

$$\mathbb{A}_i = \begin{cases} \mathbb{A}_i + A_a & (C(s, s_{\pi_i}) \geq T_a) \\ \mathbb{A}_i & (\text{Otherwise}) \end{cases}, \tag{9}$$

ここで、Aaは活性化係数値、C（s、sπi）はsとsπiの間の抽出された特徴の類似性の関数であり、sπiは方策πiがエージェントによって取得されるときの環境特徴です。

C.拡散機能

SAP-netに含まれる各方策は、活性化された値を隣接する方策に分散できます。簡略化された拡散が図4に示されています。方策πiが値Aiをアクティブにした場合、活性化された値ηiの伝播として隣接するπjに拡散します。

$$\eta_i = \frac{1}{N}\mathbb{A}_i e^{-w_{ij}}. \tag{10}$$

ここで、Nは方策πiの出力パスの数です。したがって、方策πiは、式（10）から計算された同じ活性化された値を出力します。方策πjは、図4に示すように、隣接する方策から複数の活性化された値を受け取ります。πjが受け取った活性化された値の合計は、

$$\mathbb{A}_j \leftarrow \mathbb{A}_j + \sum_{k=1}^{n} \eta_k. \tag{11}$$

たとえば、図5のように2つの拡散ターゲットが活性化場面に存在する場合、各活性化拡散値は式（4）によって計算されます（10）。方策πjとπkは、πiの活性化拡散をした後、値を相互に拡散します。方策πjはπkからの拡散値ηlを受け取り、πkはπjからの拡散値ηmを受け取りました。拡散方向と拡散活性化値は次の式で決定されます。

$$\eta_l = \begin{cases} {}^{\triangle}\eta & ({}^{\triangle}\eta > 0) \\ 0 & (\text{Otherwise}) \end{cases}, \tag{12}$$

$${}^{\triangle}\eta = \frac{1}{N-1}\mathbb{A}_k e^{-w_{jk}} - \frac{1}{N-1}\mathbb{A}_j e^{-w_{jk}}. \tag{13}$$

ここで、ηは、πjとπkからの活性化値の伝播の差です。伝搬活性化値の計算では、出力パスを受信パスとして使用されることが禁止されているため、NはN-1に変更されます。

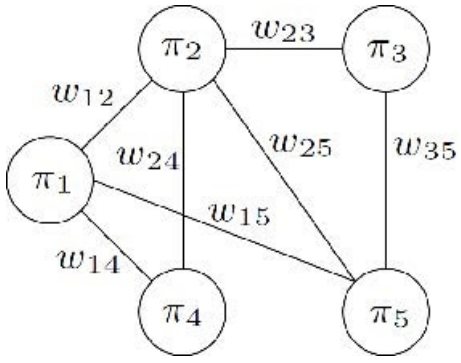

図5.活性化値の複数拡散の状況。この場合、πj とπkは同時に影響を受けます。計算結果として、活性化値は、式のようないずれかに伝播されます。（13）。

SAP-netに複数のノードが含まれていて、活性化と拡散が同時に発生した場合、SAP-netはシステムとして動作し、多体問題のようになります。このSAP-netは、基本設定として活性化の順次拡散に設定されています。

D.リコールのしきい値

再利用ポ方策を選択する前に、活性化値Aiを使用して候補方策πiが抽出されます。候補方策Πcのセットを作成するために、次のしきい値によってπiがフィルタリングされます
Aiを使用した関数T（・）。

$$\mathbb{T}(\pi_i) = \begin{cases} \pi_i \in \Pi^c & (\mathbb{A}_i > T_r) \\ \pi_i \notin \Pi^c & (\mathbb{A}_i \leqslant T_r) \end{cases}, \tag{14}$$

ここで、Πcは候補方策のセットであり、Trはしきい値です。実装では、πiはオプションの値Aciを与えます。

E.方策選択

複数の方策を同時に想起することは禁じられているため、以下の機能で方策を選択します。再利用方策は候補方策から選択されます。再利用方策は、以下によって定義されるソフトマックスのような関数によって決定されます。

$$\mathbb{S}(\pi_i) = \frac{\exp \mathbb{A}_i}{\sum_{\pi_j \in \Pi^c} \exp \mathbb{A}_j}. \tag{15}$$

再利用方策は任意のタイミングで決定されます。今後の作業では、複数の方策を同時に再利用し、それらの同化を図ります。

F.減衰プロセス

活性化された値は、各タイムステップで減衰します。時間や行動と同期します。このメカニズムは、人間の忘却現象に触発されています。長期的な学習でSAP-netの状態を再メッシュするために提案された方法に減衰プロセスが実装されています。

この研究での減衰値は、

$$\Delta\mathbb{A}_i = \begin{cases} 0 & (\mathbb{A}_i = 0) \\ d & (\mathbb{A}_i > 0) \end{cases}, \tag{16}$$

ここで、ΔAiは、πiでの活性化された値Aiの減衰値であり、dは減衰定数の値です。この減衰値は、現在のアクティブな値Aiを減らすために次の方程式で使用されます。

$$\mathbb{A}_i = \mathbb{A}_i - \Delta\mathbb{A}_i. \tag{17}$$

G.アクティベーション評価

提案手法をエージェントによる行動まで、選択方策再利用の有効性を評価します。エージェントがポジティブトランスファー（PT）効果を観察した場合、SAP-netの重みは、以前に使用された方策と現在の再利用方策の間で小さく調整されます。したがって、SAP-netではネットワークパスの長さが短くなります。ネガティブトランスファー（NT）の場合、重量も大きな値に調整されます。したがって、ネットワークパスの長さが長くなります。

この研究では、重みwijは各行動で調整されます。調整機能は以下によって定義されます。

$$w_{ij} = \begin{cases} w_{ij} - w_p & \text{if PT is emarged} \\ w_{ij} + w_n & \text{if NT is emarged} \end{cases}, \qquad (18)$$

ここで、wijは現在の再利用方策πiと以前に再利用された方策πjの間の接続の重みです。ポジティブトランスファーの場合、重みwijはwij <1を使用して制御する必要があります。

活性化された方策の活性値は、行動の結果によって評価されます。選択した方策での行動に基づいてエージェントが障害物に衝突するなど、否定的な転送が発生した場合、活性値はペナルティ値を提供します。この関数は、

$$\mathbb{A}_i = \begin{cases} \mathbb{A}_i - A_p & (\text{NT is emerged}) \\ \mathbb{A}_i & (\text{Otherwise}) \end{cases}. \qquad (19)$$

VI. 結論

この論文は強化学習における転移学習のための新しい方策選択法を提案しました。提案された手法は、認知科学で議論されている拡散活性化理論に触発されています。提案されているSAP-netには、ネットワーク化された保存済み方策、アクティベーション機能、拡散機能、減衰機能、想起機能などの機能が含まれています。基本的な実験は、学習されたポリシーを選択できる強化学習エージェントを使用して、最短経路問題で実行されました。実験結果から定量的な評価が行われ、SAP-netを備えた学習エージェントがWTおよびPT条件よりも早く問題を解決できることが結果から示唆されています。 SAP-netを備えたエージェントは、環境から適応的に方策を選択します。

将来的には、パラメータ設定について検討する必要があります。 SAP-netは活性化と減衰に敏感であり、トレードオフに関連しています。高い活性値が設定されている場合、減衰値は活性値をキャンセルできないため、上昇し続けます。さらに、減衰はSAP-netの影響を強く受け、活性値が上昇しない場合があります。提案された方法は、順次実行として構築され、計算コストは方策の数とともに増加します。提案される方法の順序はおよそO（n2）であり、計算は逐次処理です。並列化の方法も、実装フェーズの重要な問題です。さらに、動作SAP–netはN体問題に接続されているため、システム動作にはより理論的な考慮が必要です。