F-026

観測されたシンボルの関連性を用いた強化学習の転移学習

Transfer Learning of Reinforcement Learning Using Relationships of Observed Symbols

利根 義宣† 金子 貴輝† 長谷川 修†
Yoshinori Tone Takaaki Kaneko Osamu Hasegawa

1. まえがき

移動ロボットを実環境で効率よく稼働させるため，本稿では，環境から観測できるシンボル間の関連性を自律的に学習し，学習後の知識を他環境に転用して強化学習を高速化する手法を提案する．

2. 提案手法

本研究で扱う問題はエピソード的タスクとする． タスク内で観測されるシンボルはその種類を示す ID をもっており，提案手法は各シンボル ID とゴールとの関連性を重要度として算出し，学習する．学習するタスクはソースタスクとターゲットタスクからなる．両タスクでは同じシンボル ID のゴールを探索する．

2.1 ソースタスクにおけるシンボル関連性の学習

ソースタスクではエージェントはゴールまでの経路を強化学習すると同時にシンボルの関連性を学習する．エージェントには視界があり，視界に入ったシンボルの ID とゴールから逆算して何ステップ目に視界に存在したか，視認したシンボルのユニーク数を記憶する．エピソードが終了し，ゴールシンボルに到達した場合，式(1)と記憶から観測したシンボル ID の重要度を更新する．

$$\mathrm{I}_{id}(\mathrm{t}) = \mathrm{I}_{id}(\mathrm{t}-1) + \alpha_{imp}\left(R\exp\left(\lambda(\tau_{id} + \mu\, N_{id})\right)\right) \quad (1)$$

$\mathrm{I}_{id}(\mathrm{t})$はシンボル ID の重要度，R は報酬，$\tau_{id}$はエピソード中に観測したシンボル ID のゴールまでのステップ数平均，N_{id}は遭遇したシンボル ID の個数を示す．

重要度は各シンボルとゴールシンボルの関連性の強さを示す．式(1)により，ゴールシンボルに近く，また環境上における個数が少ないほど重要度は高くなる．一方，例えゴールシンボルに近いシンボルであっても環境上に多数存在するような陳腐なシンボルは重要度が低くなる．

2.2 ターゲットタスクにおける転移学習

ターゲットタスクでは学んだ重要度を転移し，重要度に基づく擬似報酬を自らに与えることで強化学習の探索にバイアスをかける．

提案手法は，Q-Learning によって学習する複数の行動価値関数を有し，状況に応じて行動選択に利用する行動価値関数を切り替えることで，重要度の高いシンボル周りを探索しながら，ゴールを目指すことを可能にしている．用いる行動価値関数は 3 つあり，疑似報酬を得て学習するシンボル探索 Q と環境からの報酬を得て学習する楽観的ゴール探索 Q，ゴール探索 Q がある．楽観的ゴール探索 Q はオプティミスティック初期値 1.0 を用いて初期化する．

エージェントはエピソード開始と共にゴール探索フェーズとなる．ゴール探索フェーズでは楽観的ゴール探索 Q とゴール探索 Q が式(2)，(3)の確率により選択され，ε-Greedy により行動を選択する．

$$\mathrm{P}pos\,(e) = 1/(1+\exp(-a(c\,e-b))) \quad (2)$$

$$P_{nega}(e) = 1 - P_{pos}(e) \quad (3)$$

e は経過したエピソード数である．

式(2)，(3)による確率の遷移により，学習序盤は楽観的ゴール探索 Q により活発な探索行動を行い，学習が進むにつれゴールシンボルに向かう行動に収束する．

ゴール探索フェーズは事前に決められたステップ数 MaxExplorerStep の期間続き，その後はシンボル探索フェーズに切り替わる．シンボル探索フェーズはシンボルが発見されるまで続き，シンボルを観測するとゴール探索フェーズに切り替わる．

シンボル探索フェーズではシンボル探索 Q を利用した ε-Greedy により行動を選択する．このフェーズ中にエージェントによってシンボルが観測されると，1 タスクにつき各シンボルごとに 1 度だけ，重要度に基づいた擬似報酬 I_{id} がシンボル探索 Q に与えられる．よって，このフェーズの間は重要度の高いシンボルへエージェントが向かいやすくなる．また，1 度擬似報酬を得たシンボルからは 2 度と報酬を得ないので，そのシンボル付近を行動するにつれ，その場付近の状態価値が環境の報酬に従って低下する．よって，重要度の高いシンボルであっても，周りに次なる有力なシンボルが無い場合，エージェントはその場を何度も訪れようとするので状態価値が低下し，次第にエージェントが向かわない状態となる．この働きにより，重要だと判断されていたシンボルが適切でなかった場合でも，そのシンボルへの固執をやめ他の探索に移ることができる．再度シンボルが観測されると探索ステップが初期化され，次ステップではゴール探索フェーズに切り替わる．

上記のアルゴリズムの働きにより，学習序盤は重要度の高いシンボルを中心とした探索を行い，終盤ではシンボルにとらわれないゴール探索へと移行する．また，転移利用した重要度の知識が転移先では適切でなかったとしても，誤ったシンボルへの固執をやめ，ゴール探索へ移行することができ，大幅な学習の遅延は回避される．

3. 実験

3.1 実験環境

提案手法の有効性を確認するため，実在する食料品店のレイアウトを用いたシミュレーションによる探索実験を行った．入り口をスタートとして，各売り場をゴールとする．この問題ではシンボルは売り場であり，シンボル ID はその種類を示す．スーパーマーケット A（地図 A）と B（地図 B）の店舗内地図は図 1 のようになっている．この地図

†東京工業大学

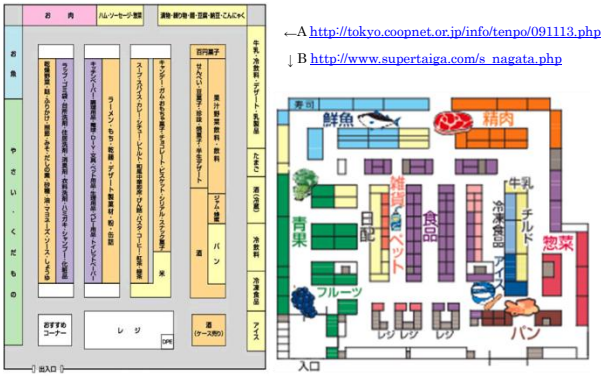

図 1:転移学習に用いる
スーパーマーケット A と B

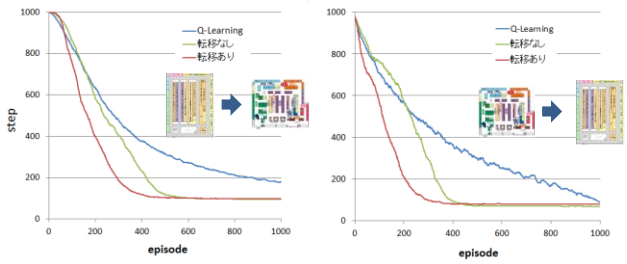

図 2: ゴールするまでの学習曲線
(左)店 A で学習し店 B でテストした結果
(右)店 B で学習し店 A でテストした結果

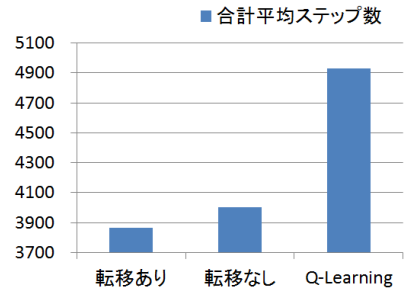

図 4:15 の売り場について,
店 A で学習し店 B でテストしたステップ数と
店 B で学習し店 A テストしたステップ数の合計

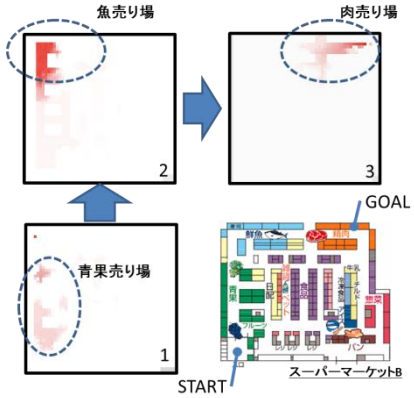

図 3:提案手法のスーパーマーケット B における
肉売り場探索の注目シンボルの遷移

をもとに強化学習の環境を生成した．ただし，エージェントがそれぞれの売り場に接近すると，売り場のシンボルを認識できるようになっている．売り場の種類は 15 あり，また，ノイズとしてランダムに地図上の 10 地点を選び出しノイズシンボルを設置した．

3.2 実験方法

実験は，地図 A と地図 B の一方をソースタスク，もう一方をターゲットタスクとする．エージェントがゴールシンボルに到達するか，1000 ステップ経過した時点で 1 つのエピソードが終了する．それを 2000 エピソード行う．報酬はゴールに到達した時に 1.0 が与えられる．ソースタスク，ターゲットタスク共に同一のシンボル ID をゴールとする．

実験における提案手法の有効性を示すため，転移なしエージェントと Q-Learning エージェントとの比較を行った．転移なしエージェントは提案手法のアルゴリズムから重要度の転移を排除し，ターゲットタスクにて全てのシンボル ID の重要度を一律 0.5 としたものである．Q-Learning エージェントは Q-Learning を行い，行動価値の初期値を 1.0 とすることで状態空間の探索効率を上げている．それぞれのエージェントにおける行動選択は ε-Greedy を使用した．

3.3 実験結果

結果は全て 50 回実験を行った平均である．代表的な売り場探索の例として肉売り場探索をした時のグラフを図 2 に示す．ゴールに必要なステップ数のグラフから，転移ありの学習が他エージェントよりも早いことが分かる．これは，過去に学んだ重要度をヒントに探索を行うことで効率的にゴールシンボルまでたどり着いている事を示す．各売り場をゴールとする各エージェントの平均ステップ数について，全体の約 7 割の売り場探索で転移手法が他手法よりも平均ステップ数が少ない結果となった．転移情報を使うことにより，最適経路をいち早く学べていることが分かる．また，図 4 は全実験のステップ数の合計だが，転移ありがゴールへの経路を早く学べている様子が再度確認できる．

図 3 は提案手法のスーパーマーケット B における肉売り場探索にて，シンボル探索 Q の行動価値の遷移を表しており，赤い箇所は高い行動価値を示している．転移元で学習された重要度の大きさの序列は肉，魚，青果であったが，この図からエージェントのシンボル探索が青果，魚，肉売り場と移動し，学んだシンボルの関連性を活かせていることがわかる．

4. むすび

本研究では，環境におけるシンボルの関連性という高次の知識を学習し，新たな環境でその知識を効率的に利用することで強化学習を高速化する転移学習手法を提案した．そして，実在する食料品店における売り場探索実験にて，他食料品店の売り場配置の関連性を学習し転移することで，店舗の構造が違っていても強化学習の高速化を行うことができることを示し，提案手法の有効性を示した．今後の課題として，PIRF-Nav[1]との組み合わせを考えることでナビゲーションのタスクに応用していきたい．

参考文献

[1]森岡博史ほか: ”人の多い混雑な環境下での SLAM による移動ロボットのナビゲーション”，第 28 回日本ロボット学会学術講演会，(2010)