論文の内容の要旨

論文題目　行動経験の変換不変性に基づく移動ロボットの行動学習

氏名　増山 岳人

本論文では，過去に獲得した知識を用いて積極的に探索空間を縮減することで，未知環境におけるロボットの試行錯誤的な行動学習を効率化する手法を提案する．提案手法は，短時間的な行動の組み合わせから構成されるスキルの獲得と利用によって学習の効率化を行う，内発的動機づけを導入した階層型強化学習を基盤としている．ここでの内発的動機づけとは，特定のタスクに依存しない形式で用いられる報酬信号による動機づけである．

従来の内発的に動機づけられた階層型強化学習では，環境とロボットとの相互作用の過程から有用なスキルを累増的に獲得し，スキル間の関係性を構造化する方法論についての議論が中心であった．しかしながら，学習の進行に伴ってより長時間的なスキルが獲得されるようになり，行動の階層構造が拡張される程探索空間が拡大し，階層構造の上位層における見かけ上の学習速度の低下が起こる．これは状態行動空間における学習の効率化や，その継続性を重視し，網羅的な経験の収集を試みる探索戦略に起因する問題であると考えられる．この問題に対し，本論文ではスキルを用いて積極的に探索空間を縮減し，選択的な探索に基づく学習を行う手法について論ずる．スキルに基づく探索空間の縮減を行うことで，既存知識の現在の環境における利用結果を評価し，探索空間を下位層の行動空間から上位層のスキル空間に移すことが本論文の狙いである．

まず，本論文ではスキルは過去の成功経験から抽出されるものとする．具体的には，ある特定のタスクに対する最適方策の実行によって得られる有限の行動の順序集合と，それらの行動系列の実行に伴って観測されるセンサ情報の系列を低次元化した量の直積集合という形式でスキルを定義する．センサ情報の系列を低次元化した量は，新たな環境におけるスキル実行の結果に対して，過去の成功経験の再現性を抽象的に評価する尺度として利用する．経験の再現性に対して内発的動機づけを与え，スキルの価値に基づいた行動選択過程へのバイアスを印加することで，スキルによって構成される経路を中心とした指向性をもった探索が行われる．さらに，そのような経路候補はタスクを記述する外発的報酬によって絞りこまれ，状態空間を網羅的に探索することなく，効率的に学習を行うことが可能となる．

経験の再現性の尺度となる時系列の低次元化量としては，アファイン変換不変量を導入する．これにより，特徴空間においてアファイン変換でモデル化される不要な情報に対する不変性を利用して，抽象的な経験間の対比を行う．本論文では，移動ロボットのナビゲーション問題において，観測情報の対称性などの情報を捨象するためにアファイン変換不変量を利用する．これを用いて，成功経験の再現性を特定の観測情報からの距離などではなく，抽象的な形式で計る．

具体的な学習手法は，Q-learning などの TD 学習（Temporal Difference learning）を基盤としている．TD 学習などで一般的に用いられる行動価値関数に加えて，ある状態におけるスキルの価値を表すスキル価値関数を並列に学習する形式となっている．ロボットの行動は，各状態における行動価値関数に基づいて選択される．他方，各状態ではいずれかのスキルが選択されており，行動選択過程においては選択されているスキルに基づく一時的なバイアスが，スキルの指定する行動の価値に加えられる．スキル価値の学習過程には外発的報酬だけでなく，スキル実行の結果得られる観測情報から計られる，経験の再現性に対する内発的動機づけが与えられる．また，スキル価値の更新にはスキル実行の結果遷移する状態における行動価値が利用される．そのため，スキル実行の結果，よりよくタスクを実行し，過去の成功経験の再現性が高く，より高い行動価値をもつ状態への遷移を実現するスキル程高い価値をもつことになる．

上述の更新則にしたがってスキル価値を学習し，行動選択過程への一時的なバイアスを印加することで，高い正の価値をもつスキルが選択された場合は，そのスキルが指定する行動が選択される確率が上昇する．逆に負の価値をもつスキルが選択された場合は，そのスキルが指定する行動が選択される確率は低下する．その結果，スキルとその価値によって行動選択の戦略に偏りが生じ，探索空間が縮減することになる．この探索空間の縮減によって以下のような効果を得ることができる．

学習初期においては，全ての状態において行動及びスキル価値はほぼ同一の値をとっている．そのため，Q-learning や Sarsa といった代表的な TD 学習手法では初期状態を中心として等方的に探索範囲を広げる探索戦略がとられる．他方，提案手法ではスキル価値によって探索が方向づけられる．提案手法では，状態行動空間における探索と同様に，状態スキル空間における探索が行われるが，スキル価値の更新則には過去の成功経験に対する再現性に対する内発的報酬が与えられる．経験の再現を試みること自体はタスクの実行効率に直接寄与するものではないが，これによって探索は等方的ではなく，成功経験に由来する指向性をもつことになる．その結果，成功経験がもつ行動系列の指向性を担保する，探索の中心となる経路候補が行動価値上に構成される．さらに，学習が進むとそれら複数の経路候補は外発的報酬によって絞りこまれ，より選択的な探索が実行されるようになる．そしてスキルの指定する行動の，タスク実行という目的に対する整合性は，外発的報酬に基づく行動価値の学習によって調整される．その結果，通常の TD 学習と異なり，提案手法では全ての状態行動対を訪問するような探索戦略はとられず，スキルという既存の知識に基づいた選択的な学習が行われる．これにより，学習速度を向上させることが可能となる．

以上の提案手法の有用性を，本論文では２次元グリッドワールドにおけるナビゲーション問題を例に検証している．過去の成功経験に基づいた指向性をもった探索を行うことの効果として，学習初期において行動価値とスキル価値の並列学習構造によって学習時間が大幅に短縮されることが示されている．また，経験の再現性に対する内発的報酬の導入によって，収束性能が向上するという結果が示されている．さらに，提案手法の探索空間の縮減効果によって，状態数が増大しても安定した学習性能の向上効果を得ることができることが示されている．

上述の提案手法により，タスクに対して適切なスキルに基づく探索空間の縮減効果が学習を加速することが可能となる．しかしながら，累増的なスキル獲得を想定した場合には，ロボットが未知環境において

探索に利用できるスキルは必ずしもタスク実行において有用なものばかりではない．また，環境と身体の複雑さに応じて，スキル数は膨大なものになる場合があると考えられる．タスクに対して不適切なスキルの実行と，スキル数増大による一回のスキル実行当たりの相対的な学習量の低下は，ともに学習効率を低下させる要因となる．本論文では，この問題に対し，スキルの類似度に基づく適格度トレースを導入している．一般的な適格度トレースでは，実際に実行した行動やスキルのみに対して高い適格度を付与する．本論文では実際に実行していない全てのスキルについても，実際に実行されたスキルに対する類似度に応じた適格度を与える．これにより，スキルが多様化し，スキル数が増大した設定においても有用な探索空間の縮減を行うことが可能となる．

未学習の環境において，不適切なスキルの実行はスキル価値を低下させる．類似度に基づく適格度トレースによって，斉次的にスキル価値を更新することで，スキル実行当たりの相対的な学習量低下の問題が解決し，負のスキル価値に基づく探索空間の縮減が利用可能となる．さらに，変換不変性を用いた経験の再現性の評価に基づく内発的異報酬により，探索中心の候補となる経路が行動価値関数上に構成される．その結果，様々な方向性をもった，大量のスキルを利用する場合にも，提案手法による探索空間の縮減効果を得ることが可能となる．

以上のように本論文では，スキルという知識に基づく探索空間の縮減機能を強化学習手法に実装する枠組みを提案し，その有用性を示している．未知環境で自律的に運用可能なロボットシステムの構築のためには，ロボットのセンソリモータ系の，環境に応じた適応的な構造化が必要となる．本論文において示された結果は，そのような手法がもつ探索空間の拡大に伴う学習時間の増大という問題に対し，既存の知識を利用することによる選択的な学習の重要性を示唆している．