課題番号	Q22D-06
課題名（和文）	タスク全体の大局的な基準値を用いた満足化による強化学習の探索率制御
課題名（英文）	Search Rate Control by Satisficing using Whole-Task Global Reference Value in Reinforcement Learning Task
研究代表者	所属（学部、学科・学系・系列、職位） 東京電機大学　先端科学技術研究科　情報学専攻 宝田　悠
共同研究者	所属（学部、学科・学系・系列、職位） 東京電機大学 理工学部理工学研究科　情報学専攻 武井　介
	所属（学部、学科・学系・系列、職位） 理工学部　理工学科　教授 高橋　達二
	所属（学部、学科・学系・系列、職位） 氏名

研究成果の概要（和文）

強化学習は機械学習の枠組みの一種であり、行動主体(エージェント)と環境の相互作用によって学習を行う。強化学習では、エージェントが探索を行い情報収集するか学習した情報から報酬獲得をするかを自律的に制御する必要があり、探索と活用のトレードオフとして強化学習タスクにおける解決すべき根本的な課題として扱われている。この問題の解決策の一つとして、本研究では S2E 法を提唱した。S2E 法で算出した「満足値」と呼称する値と探索率の変換方法に関して、一部分の計算に sigmoid 関数を用いることによって、他の代表的なアルゴリズムと比較してより性能が向上していることが確認できた。

研究成果の概要（英文）

Reinforcement learning is a part of the machine learning framework, in which learned behavior (learning) occurs through the interaction between the agent (the subject of the action) and the environment. In reinforcement learning, the agent requires autonomous control over whether to search and collect information or to earn rewards from information it has learned, and this trade-off between search and exploitation is discussed as a problem that must be solved as a fundamental issue in reinforcement learning tasks. As one of the solutions to this problem, we proposed the S2E method, which uses the sigmoid function to calculate a part of the conversion between the "satisfaction value" calculated by the S2E method and the search rate, and we confirmed that the performance was improved compared to other typical algorithms.

1. 研究開始当初の背景

近年の機械学習における技術的発展によって、機械学習技術は様々な分野で実務的な利用をされている。機械学習の手法の一つとして強化学習がある。強化学習の特徴は、対象の環境で自律的に行動し、行動した結果として環境から受け取るフィードバックを用いることによる、自律的な学習にある。この強化学習に関して、議論すべき問題として探索と活用のトレードオフに関する問題がある。学習に必要な探索と学習結果を利用した活用は同時に行うことができない。十分な学習が行えていない状態では、学習結果を利用しても効率的な行動は行えず、強化学習エージェントが十分に学習を行った後に活用する行動をとる必要がある。そのため、適切なタイミングで探索と活用を切り替える為の手法を考えることが必要である。例えば探索範囲が広いなどで探索が大量に必要な環境では、環境を完全に学習し把握するのに必要な探索量が膨大になり、行動を最適化することは困難である。しかしながらこの問題に関しての決定的な解決手法はいまだに確立されていない。

2. 研究の目的

本研究では、人間の認知傾向の一つである満足化に着目した。人間は広大な環境で探索を行う際に、最適解を見つけるのではなく、目的として基準を設定し基準を満たす満足解を探索する傾向がある。Simon らはこの傾向を満足化と呼んだ。本研究では、強化学習における探索と活用のトレードオフ問題へのアプローチとして満足化を用いたS2E法を提唱する。S2E法では、満足化を強化学習の探索率の調整として適用している。

本研究の目的は、強化学習の枠組みに人間の認知機能の一つである満足化を適用させる手法として S2E 法を提唱し、その挙動や性質を検証することである。

3. 研究の方法

本本研究で提唱するS2E法では、スタートからゴールまでのエージェントの行動から、満足度と呼称する数値を算出し、満足度を探索率に変換を行う。その際に行う変換方法に関して模索を行なった。満足度は今回の場合、どれだけ最短距離に近いルートを通ることができたかを示している。単純に満足度を探索率の範囲に正規化した S2E と、sigmoid 関数を用いた変換を行う S2E、また最終的に一部を線形、一部にsigmoid関数を用いた変換するS2E-sigも含めて検討を行なった。この変更には、満足度の高低による探索率の非線形的な関連性を表現する狙いがあった。

アルゴリズムの性能比較は、コンピュータ上でのシミュレーションによる実験を用いて行った。実験は、GridWorldタスクと呼ばれる迷路タスクで実験を行った。GridWorldタスクは、マス目状のマップで経路探索を行うタスクである。このタスクを用いてスタート位置からゴール位置までの最短経路を学習する速度を比較する上で、指針としてスタートからゴールまでのステップ数を用いた。

4. 研究成果

GridWorld タスクでのステップ数遷移を代表的なアルゴリズムとの比較を行なった結果を図 1 に示す。

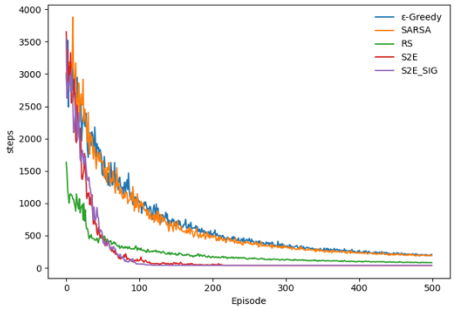

図 1 GridWorld タスク(20×20)における
各アルゴリズムのステップ数推移

図 1 の比較対象アルゴリズムのなかで S2E-sig が最も早いタイミングで収束していることがわかる。このことより、S2E-sig が素早く最適解への収束を行うという側面での効率的探索を行うアルゴリズムとして有用であるということが言える。また、同じ満足化の枠組みを取り入れている RS との比較をすると、各アルゴリズムの特徴を見ることができる。RS アルゴリズムは探索初期段階において素早く探索を行い、ステップ数をいち早く削減することができている。しかし一定の満足解に達してしまうと探索効率が下がっている。一方で S2E-sig は、探索初期では RS ほどの素早い探索は見られないが最適に近い解を求める速度に関しては RS よりも優れていることが示唆されている。

本研究は S2E-sig の検証を通じて、強化学習の重大な課題である「探索と活用のトレードオフ」に対する一つの方法を提示した。活用促進の判断を自律的制御が目下の課題である。