氏　　名　奥戸 嵩登

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2414 号

学位授与の日付　2023 年 3 月 24 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Reward Shaping with Human Subgoals

論 文 審 査 委 員　主　査　山田 誠二
情報学専攻 教授
杉山 麿人
情報学専攻 准教授
稲邑 哲也
情報学専攻 准教授
荒井 幸代
千葉大学 大学院工学研究院 教授
山口 智浩
奈良工業高等専門学校 情報工学科 教授

（様式３）

博士論文の要旨

氏　　名　奥戸　嵩登

論文題目　Reward Shaping with Human Subgoals

Many researchers have actively studied Reinforcement learning, which acquires a policy maximizing long-term rewards. Unfortunately, this learning type needs to be faster and easier to use in practical situations because the state-action space becomes enormous in the real world. Many studies have incorporated human knowledge into reinforcement Learning. Human knowledge of trajectories is common, but it could ask a human to control an AI agent. Controlling the AI agent could be too hard in specific tasks like robotics. Knowledge of subgoals may lessen this requirement because humans only need to consider a few representative states on an optimal trajectory. The essential factor for learning efficiency is rewards. Potential-based reward shaping is a primary method for enriching rewards and realizes a policy-invariant reward transformation which remains the optimal policy for an original reward function. A potential function is a real-valued function given a state, and the difference between its output in the current state and one in the previous state becomes a shaping reward. However, incorporating subgoals for accelerating learning over potential-based reward shaping is often challenging because the appropriate potentials are not intuitive for humans. We propose *subgoal-based reward shaping* based on potential-based reward shaping. Subgoal-based reward shaping includes a potential function given state history and time. We prove that subgoal-based reward shaping is policy-invariant.

Subgoal-based reward shaping makes it easier for human trainers to share their knowledge of subgoals. Since the potential function is essential to make learning efficient, we proposed two types of the potential function, *static goal-oriented potential* and *learned potential* in subgoal-based reward shaping. Static goal-oriented potential approximates an optimal value function because the current study indicated that potential-based reward shaping made policy learning efficient when the potential function was the optimal value function. We define a hyperparameter of static goal-oriented potential, which controls the shape of the potential. The evaluation result indicates that the hyperparameter deteriorates learning efficiency when inappropriate. To solve this challenge, learned potential acquires its potential simultaneously with policy learning to remove the hyperparameter. We adopt a value function over abstract states, which updates with n-step temporal difference~(TD) learning during policy learning, as a potential function in learned potential. The abstract state is a subgoal achievement and begins from underachievement, and the transition of the abstract state follows the order of subgoals.

We conducted a user study to collect subgoal sequences from participants. The subgoals acquired from participants are more biased than the random-generated subgoals, and many participants provide the same or similar subgoals.

We conducted simulation experiments in three domains covering discrete and continuous states and actions. The experimental results indicate the effectiveness that subgoal-based reward shaping makes several baseline reinforcement learning algorithms, including a deep reinforcement learning algorithm, efficient. Learned potential achieves similar performance to static goal-oriented potential. The results also indicate that the participants' subgoal sequences are superior to the random-generated subgoal sequences for subgoal-based reward shaping.

The performance analysis between participants' and random subgoal sequences shows that the performances in the domains where the subgoals are on optimal trajectories are similar. The best subgoal sequence of the participants is a part of more optimal trajectories than others. We found that an appropriate number of subgoals and a subgoal on an optimal trajectory can improve the baseline algorithm. Subgoal-based reward shaping performs well with a partially ordered subgoal sequence. Static goal-oriented potential is sensitive to the change of the hyperparameter, and initializing the potential improves the performance of learned potential. Subgoal-based reward shaping cannot improve the baseline algorithm in negative step rewards. Learned potential is better than static goal-oriented potential in mixed positive and negative rewards.

This dissertation does not propose an easy way to collect subgoal sequences from humans, but subgoal-based reward shaping can be applied to many domains as long as the user has subgoals. Though negative step rewards disable the effectiveness of subgoal-based reward shaping, they can convert to a positive goal reward which subgoal-based reward shaping works well. Subgoal-based reward shaping can improve baseline reinforcement learning algorithms when a subgoal sequence is on an optimal trajectory. A detailed methodology might be optional, and it is enough for a subgoal teacher to have an optimal trajectory and subgoal sequence. This dissertation might encourage people to use subgoals yet to be used and help accelerate reinforcement learning.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　奥戸　嵩登

Title
論文題目　　Reward Shaping with Human Subgoals

本学位論文は，「Reward Shaping with Human Subgoals」と題し，全８章から構成されている．

第１章「Introduction」では，研究の動機，人工知能，機械学習研究への貢献，研究目的そして本論文の概要について説明されている．次に，第２章「Related Work」では，強化学習に人間の知識を導入する研究，サブゴール知識を強化学習に活用する研究，リワードシェイピングなどの関連研究について説明がされている．第３章「Background」では，強化学習の枠組み，様々な強化学習アルゴリズム，ポテンシャルベースのリワードシェイピング，動的リワードシェイピング，そしてサブゴールの概念などの研究背景について説明がなされている．ここで，本研究の目的が「サブゴール知識を利用可能にしたリワードシェイピング手法による，サブゴール知識を使った（深層）強化学習の高速化」であることが説明された．続く第４章「Subgoal-based Reward Shaping」において，人間から与えられたサブゴールを有効利用するための強化学習アルゴリズムの提案について説明されている．提案アルゴリズムは，リワードシェイピングの枠組みにサブゴールでの報酬を単調増加するように固定して与え，サブゴール間の状態遷移を抽象状態として扱うことで強化学習の高速化を実現する．さらに，その固定されたサブゴール報酬を動的に学習していく方法に拡張している．そして，第５章「Collecting Subgoal Sequence Provided by Human」では，グラフィカルユーザインタフェースを用いて人間からサブゴールを収集する方法が述べられている．続く第６章「Experiment for Evaluation」では，グリッドワールド，２次元連続空間，３次元連続空間などの様々なタスクにおける評価実験の計画，実施，そして実験結果の分析について説明されている．その結果，従来法に対する提案手法の有効性が複数のタスクにおいて確認された．最後に，研究全体の考察（第７章「General Discussion」）と結論（第８章「Conclusion」）が述べられている．

公開発表会では博士論文の章立てに沿って発表が行われ，その後に行われた論文審査会及び口述試験では，審査員からの質疑に対して適切に回答がなされた．

質疑応答後に審査委員会を開催し，審査委員で議論を行った．審査委員会では，出願者の博士研究が「人間の知識であるサブゴール知識を強化学習に利用するアルゴリズムの提案および多角的な実験的評価を行った新しい試み」であることが評価された．

以上を要するに本学位論文は，人間のもつサブゴール知識により，強化学習の高速化を実現したオリジナリティの高い研究であり，人工知能，機械学習の研究分野の発展に貢献するという点で学術的価値が大きい．また，本学位論文の成果は，学術雑誌論文１編，査読付き国際会議論文 １ 編として発表され，社会的な評価も得ている．以上の理由により，審査委員会は，本学位論文が学位の授与に値すると判断した．