内発的動機付けと対照学習の改良によるスパースな報酬環境における探索効率向上

Enhancing Exploration Efficiency in Sparse Reward Environments through Refined Intrinsic Motivation and Contrastive Learning

相良博喜 *1　屋藤翔麻 *2　草場壽一 *3
Hiroki Sagara　Shoma Yato　Shuichi Kusaba

*1九州大学 Kyushu University
*2東京工芸大学 Tokyo Polytechnic University
*3東京大学 University of Tokyo

We propose a method to enhance exploration efficiency in reinforcement learning environments with sparse rewards by improving intrinsic motivation and contrastive learning. Intrinsic motivation is used to guide agents to novel states by providing internal rewards based on prediction errors. However, existing methods such as Random Network Distillation (RND) suffer from declining intrinsic rewards over time, leading to limited exploration in later stages. To address this, we introduce two modifications: (1) adjusting intrinsic rewards when episodes terminate due to agent failure, and (2) enhancing contrastive learning by treating temporally adjacent frames as positive examples. These modifications improve novelty detection and encourage consistent exploration. Experiments in sparse-reward environments such as Climber demonstrate significant improvements, with our method achieving external rewards in half the training steps compared to baseline models. Our approach maintains high intrinsic reward peaks throughout training, paving the way for more effective exploration in challenging environments.

1. 研究背景・目的

強化学習 (Reinforcement Learning; RL) の分野は近年目覚ましい進歩を遂げている。多様なタスクへの対応、意思決定から複雑なゲームの習得に至るまで、その応用範囲を広げてきた。これらの進展は、計算能力の向上と機械学習手法、特にディープラーニングのブレークスルーによって推進されてきた。しかし、報酬信号が希薄な環境における効率的な探索は依然として大きな課題である。このような環境では、エージェントが複雑な状態空間を効率的にナビゲートし、限られたフィードバックから学習する能力が求められる。この問題に対する有望な解決策として、内発的動機 (Intrinsic Motivation; IM)[?] が挙げられる。IM は、新奇性や驚きなどの内部信号を活用して探索を導き、学習効率を高める。

この分野で注目すべき研究の一つが、Random Network Distillation(RND)[?] である。RND は、新奇性検出に基づく手法であり、固定されたランダムなターゲットネットワークと予測モデルを用い、予測モデルがターゲットネットワークの出力を模倣しようとする際に生じる誤差を新奇性の指標として用い、エージェントに探索の指標となる内発的報酬を提供する。この手法は報酬が希薄な環境でも効率よく学習することが可能であったが、固定されたターゲットネットワークに依存するため、学習が進むにつれて誤差が少なくなり内発的報酬が消失するという課題がある。

こうした課題を克服するために提案されたのが、Self-supervised Network Distillation(SND)[?] である。SND は、ターゲットネットワークを自己教師あり学習によって動的に学習させることで、ターゲットモデルの表現を適応的に進化させ、新奇性検出の精度を向上させる。

本研究では、この SND フレームワークをさらに改良するために 2 つの取り組みを行った。最初の取り組みは、内発的動機に基づく行動において、エピソード終了時の挙動を調整するための η の最適化である。内発的動機に基づく行動は、新たな状態に対して無作為に取り組むため、短期間で敵と衝突するなどしてエピソードが終了してしまい、効率的な学習が妨げられる可能性がある。これに対処するため、高い内発的動機によってエピソードが終了した場合、その内発的動機に基づく報酬にマイナスの補正を加える方法を試みた。しかし、このアプローチでは、学習に要するステップ数の短縮やパフォーマンスの顕著な向上には繋がらなかった。これにより、次の改良に焦点を移した。

2 つ目の取り組みは、内発的動機の学習方法を見直すことである。特に対照学習 (Contrastive Learning) を行う際、Noisy TV 問題が発生することを考慮した。この問題は、エージェントが一時的な新奇性に過度に引き付けられ、環境の本質的な目標から注意が逸れる現象を指す。既往研究においても、Procgen ベンチマークの Pitfall のようなゲームで、NPC の点滅やランダム動作にエージェントの関心が不適切に引きつけられ、報酬を得ることができないという課題があった。他の環境においても、初期状態におけるキャラクターの無作為な前後移動がノイズとして作用し、エージェントの探索効率を妨げる可能性がある。この問題に対処するため、対照学習を行う際に前後フレームを考慮し、連続するフレームを同一の状態として学習させる手法を導入した。この方法により、初期状態でのランダムな動作や視覚的なノイズをエージェントの興味の対象から外し、より本質的な違いに基づく状態に関心を移すことが可能となった。このアプローチにより、より短期間で報酬を得ることができ、探索効率を向上させた。

連絡先: 相良博喜，九州大学 数理学府，福岡県福岡市西区元岡 744，sagara.hiroki.043@s.kyushu-u.ac.jp

2. 関連研究

3. 提案手法

3.1 SND の基本的な手法

SND は、内発的動機付けを活用し、報酬が希薄な環境における探索効率を向上させる手法である。具体的には、強化学習における報酬に、内発的報酬 r^{intr} と外的報酬 r^{ext} の合成を用いて、新奇性の高い行動へエージェントを誘導する。

$$r_t = r_t^{\text{ext}} + \eta \cdot r_t^{\text{intr}}$$

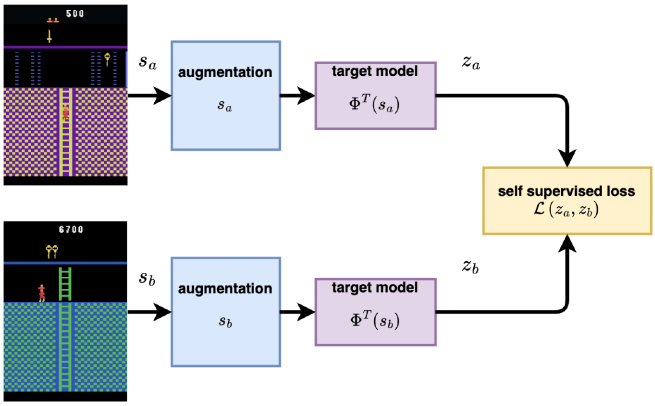

図 1: 自己教師ありネットワーク蒸留（SND）の原理。本手法は 2 つの主要な部分で構成されている。上図: ターゲットモデルを自己教師あり学習にて適切な特徴ベクトルを学習する。下図: ターゲットモデルと予測モデルの誤差から内発的報酬を得る。本図は [?] より転載。

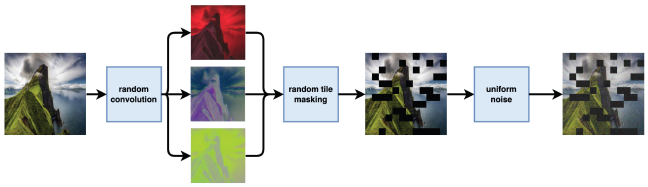

図 2: SND-V のデータオーグメンテーション方法 本図は [?] より転載。

ここで、η は内発的報酬の重み付け係数である。

内発的報酬は、予測モデル Φ_P とターゲットモデル Φ_T の間の誤差を内発的報酬として利用する [図??]。この誤差は以下の式で定義される：

$$r_t^{\text{intr}} = \|\Phi_T(s_t) - \Phi_P(s_t)\|_2^2,$$

ここで、s_t は現在の状態を表し、$\|\cdot\|_2^2$ は L2 ノルムの 2 乗を意味する。この誤差は、新奇性を評価する指標として利用され、新しい状態ほど大きな値を取る。予測モデルは、以下の損失関数で定義される通り、ターゲットモデルの出力を模倣するよう学習が行われる。

$$L_P = \frac{1}{N}\sum_{s\in S}\|\Phi_T(s) - \Phi_P(s)\|_2^2,$$

つまり、RND のようにターゲットモデルが固定されたアルゴリズムの場合、最終的には内発的報酬が収束してしまい内発的報酬が発生しなくなる。SND では、自己教師あり学習を利用してターゲットモデルを動的に更新し、状態表現の分散性を向上させる。本研究では既往論文にて提案されている SND-V というアルゴリズムを用いた。

3.2 SND-V

SND-V は対照学習をベースとした新奇性検出手法である。類似する状態の特徴を近づけ、異なる状態を分離する損失関数を用いることで、特徴空間を効果的に構築する。また、状態にノイズを付与することで、ロバストな特徴表現を学習することができる。

SND-V における損失関数の定義は以下の通りである：

$$L_T = \sum_n (\tau_n - \|Z_n - Z'_n\|_2^2)^2,$$

ただし、$\tau_n = 0$ の場合、Z_n および Z'_n は同じ状態に対してノイズを付与した状態に対する特徴ベクトルを表す。$\tau_n = 1$ の場合、Z_n および Z'_n は全く異なる状態、異なるエピソード、異なるステップに対してノイズを付与した状態に対する特徴ベクトルである。

ノイズの付与方法は既往研究と同様に [図??]、下記 3 つのフィルタを p=0.5 の確率で適用する。

・ 一様ランダムノイズ：ピクセル値に $[-1, 1]$ の範囲でノイズを追加。

・ ランダムマスキング：画像タイルをランダムにマスクする（タイルサイズは 2、4、8、12、16 ピクセル）。

・ ランダム畳み込みフィルタ

3.3 提案する改良手法

本研究では、SND の基本的な枠組みに加え、探索効率および報酬取得の向上を目指して以下の 2 つの改良手法を提案する。

3.4 改善手法１：エピソード終了時の報酬調整

内発的動機に基づく行動は、新たな状態に対して無作為に取り組むため、短期間で敵と衝突するなどしてエピソードが終了してしまい、効率的な学習が妨げられる可能性がある。これに対処するため、高い内発的動機によってエピソードが終了した場合、その内発的動機に基づく報酬にマイナスの補正を加える方法を試みた。エピソード終了時における新たなルールを導入した。具体的には、エピソード終了時に level_complete = false かつゲーム終了までのステップ数に到達していない場合、その状態を”死亡”とみなす。この場合、報酬は以下のように定義される：

$$r_t = \begin{cases} r_{\text{ext},t} - \eta \cdot r_{\text{intr},t}, & \text{死亡の場合,} \\ r_{\text{ext},t} + \eta \cdot r_{\text{intr},t}, & \text{それ以外の場合.} \end{cases}$$

内発的報酬が大きい状況にて死亡が発生した場合、その行動がエージェントの継続的な探索を妨げてしまっていると考えられる。このため、死亡時に内発的報酬にマイナスの補正を加えることで、エージェントがより長期的に探索を行うことができ、効率的な学習が可能となることが期待される。

3.5 類似状態ペア選択方法の改善

前章で述べた通り、Noisy TV 問題とはエージェントが時系列的なノイズに不適切に関心を持ち、探索が非効率化する現象を指す。この問題を解決するため、SND-V に下記の改良を加えた。

1. 同一環境および STEP が $-N$ から $+N$ ステップの範囲内にある $2N + 1$ 個の状態からランダムに選択する。
2. 選択した状態にノイズを加え、類似状態ペアを構成する。
3. 非類似状態の選択は既往研究と同様にランダムに行う。

このアプローチでは、既往研究の SND-V が採用していた完全に同一の状態ペアからノイズを加える方式よりも多様な状態表現を学習できる。特に周期的に変化する状態を同一状態と捉え、無駄な内発的報酬の発生を抑えることが期待される。これにより、状態空間全体でより精度の高い新奇性検出が可能となり、エージェントが短期的なノイズではなく本質的な環境変化に基づいて行動できるようになる。

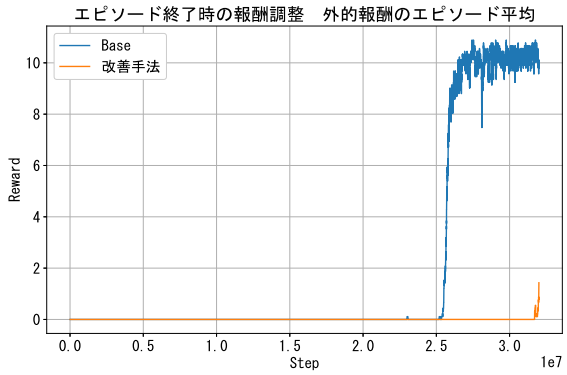

図 3: 改善手法 1: エピソード終了時の報酬調整　外的報酬の推移

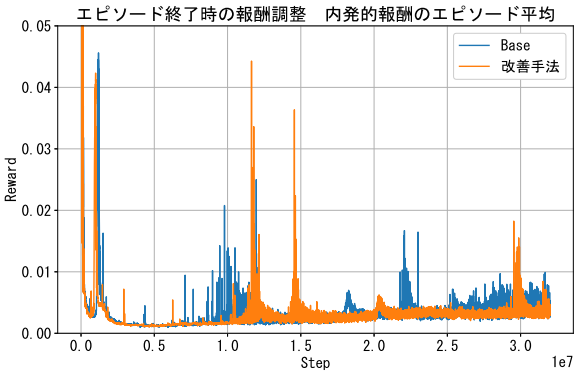

図 4: 改善手法 1: エピソード終了時の報酬調整　外的報酬の推移

4.　実験・考察

4.1　実験設定

本研究では、既往研究 [x] の条件設定を引き継ぎ、提案手法の有効性を検証した。具体的には、ProcGen ベンチマークの Climber 環境において、提案手法を適用し、探索効率および報酬取得の改善を評価した。

4.2　改善手法 1: エピソード終了時の報酬調整

まず、エピソード終了時の報酬調整の効果を検証した。[図 3]

結果から、提案手法が既往研究のモデルより外部報酬を得るまでの時間が長くなっていることがわかる。また、内発的報酬に関しても、既往研究のモデルよりもピークの発生回数が少なく、内発的動機による行動が発生していないことがわかる。このような結果となった原因を分析するため、エピソード終了時までのステップ数を生存時間解析（Kaplan-Meier 法）により評価した。[図??] まだ十分に学習が進んでいない序盤に関しては、提案手法が既往研究のモデルよりも生存時間が長い。しかし、学習が進むにつれて提案手法の生存時間が短くなっていることがわかる。これは、提案手法により序盤の探索が積極的に行われず、中盤以降の学習が遅れていることを示唆していると考えられる。

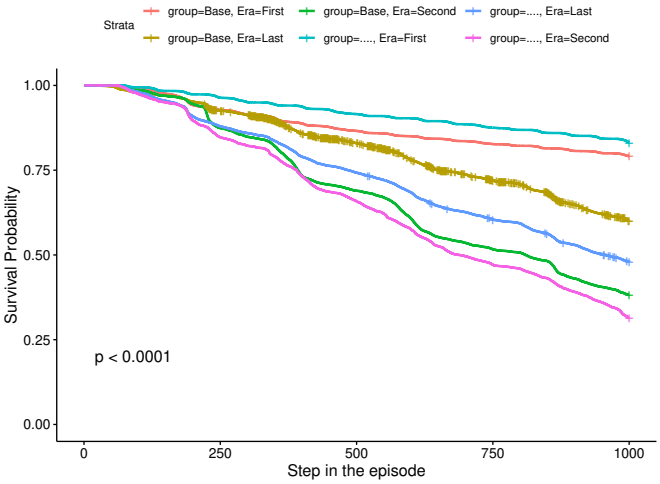

図 5: エピソード終了時までのステップ数の Kaplan-meier plot Era:First= 1 1.0e7 Step, Second=1.0e7 2.0e7, Last=2.0e7

4.3　改善手法 2: 類似状態ペア選択方法の改善

次に、類似状態ペア選択方法の改善の効果を検証した。本実験では、ペアの選択範囲は N=1、つまり前後ステップの状態を選択する方法を採用した。 [図??] が本手法の外部報酬の

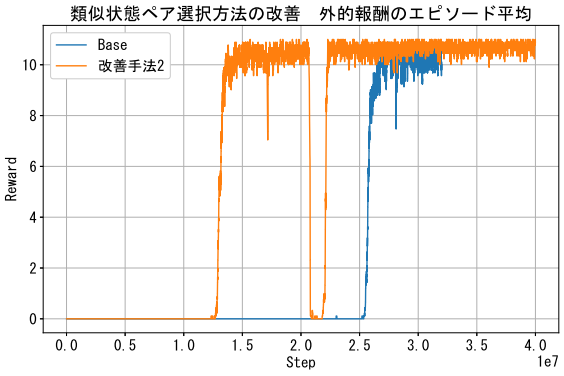

図 6: 改善手法 2: 類似状態ペア選択方法の改善　外的報酬の推移

推移である。既往研究では、2.6e7 ステップで初めて外部報酬を得たが、提案手法では 1.3e7 ステップ、つまり半分以下のステップ数で外部報酬を獲得していることがわかる。また、内発的報酬 [図??] に関しても、違いがある。0.1 0.8e7 ステップの間において、既往研究では内発的動機のピークは現れないが、提案手法では内発的動機のピークが現れていることがわかる。

学習序盤で高い内発的動機を得ることができた要因として、エージェントが適切な行動を選択することができず摂動状態に陥ることがあるが、提案手法では大きな状態変化を得ることができ、内発的動機が発生しやすくなっていることが考えられる。これにより学習序盤の探索効率が向上し、外部報酬を得るまでの時間が短くなっていると考えられる。

5.　まとめ

本研究では、内発的報酬調整と対照学習の改良により、報酬が希薄な環境での探索効率を向上させる手法を提案し、Climber

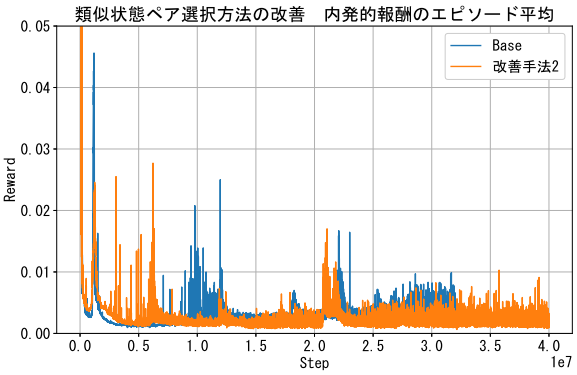

図 7: 改善手法 2: 類似状態ペア選択方法の改善　内発的報酬の推移

での実験においてその効果を実証した。今後は、より難易度の高いゲーム環境 (Montezuma's Revenge, Pitfall など) での検証を進め、提案手法の汎用性を評価する必要がある。

また、将来的な研究の方向性として、外発的報酬のスケーリングパラメータの最適化がある。

時間に基づくスケジューリング: 学習の進行に伴い、内発的動機に基づく探索を減少させることで、探索から活用へのスムーズな移行を実現する。

$$\eta(t) = \eta_{\text{init}} \cdot e^{-\alpha t}$$

このアプローチは学習初期の探索強化と後半の目標指向行動を両立できる。

報酬に基づくスケジューリング: 外部報酬の頻度や大きさに応じた適応的な探索と活用の調整を可能にする。

$$\eta = \frac{1}{1 + \beta \cdot r_{\text{ext}}}$$

報酬がスパースな環境や動的に変化する環境において、より柔軟な学習を実現する。これらの手法は、探索と活用のバランスを最適化し、効率的な学習を実現することが期待される。

参考文献

[Oudeyer 2007] P. -Y. Oudeyer, F. Kaplan and V. V. Hafner, Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Transactions on Evolutionary Computation, vol. 11, no. 2, pp. 265-286, April 2007.

[Burda 2019] Yuri Burda and Harrison Edwards and Amos Storkey and Oleg Klimov, Exploration by random network distillation, International Conference on Learning Representations, 2019.

[Pecháč 2024] Matej Pecháč, Michal Chovanec, Igor Farkaš, Self-supervised network distillation: An effective approach to exploration in sparse reward environments, Neurocomputing, Volume 599, 2024.