5P-03

PCGRL による迷路の経路形状制御手法

星野貴彦[†1]　三宅陽一郎[†2]

立教大学大学院人工知能科学研究科[†1,2]

1. はじめに

本報告では，迷路の経路形状制御問題を提案し，PCGRL（Procedural Content Generation via Reinforcement Learning）によって迷路の経路を渦巻型に生成する制御手法を検討する．

PCGRL は，コンテンツ生成に強化学習を適用するため，学習済みモデルは高速であることが見込められ，応用面で優れていると考えられる．本問題では，単に迷路を生成するだけではなく，最短経路の形状を制御することで，いかに強化学習によるコンテンツ生成が柔軟かつ強力な手法であるか調査する．渦巻型の迷路は単純かつ目標との差を認識しやすいと考え，本問題の指標とした．本研究では，迷路の経路形状制御問題を制作し，PCGRL のフレームワークを使用して学習を行い，経路の形状に対して報酬を与え，学習時に報酬がどれ程得られているか考察する．

基本的なセットアップは，迷路を生成するエージェントを作成する．このエージェントはグリッド内に壁を置く，或いは消すことができる．壁の置き場によって迷路の形状が決定されていくが，この形状を評価し，強化学習の報酬とする．図 1 に迷路の生成と学習手順を簡易的に示す．

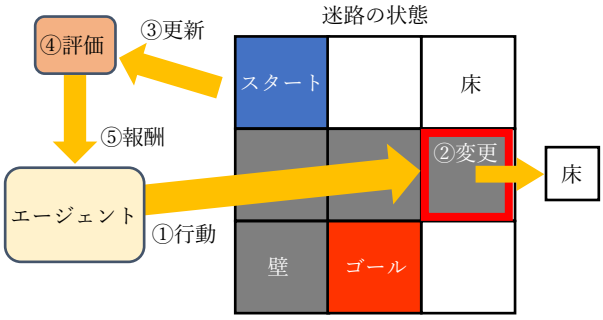

図 1 迷路の生成と学習手順

2. 関連研究

PCG は主に，構成的 PCG，探索ベース PCG に種別される．構成的 PCG は，規則に基づいて一度の過程でコンテンツを生成する手法である．探索ベース PCG は，強化学習，進化的探索，勾配に基づく最適化等の探索手法を用いて，繰り返しコンテンツに変更を加えて評価することでコンテンツ生成を行う手法である．近年では，PCG に機械学習を用いる，PCGML（Procedural Content Generation via Machine Learning）と呼ばれる手法が着目されている．PCGML には，GAN を用いた地形生成等が該当する．PCGRL も PCGML の一種である．PCG の区分や詳細な説明は[1]でなされている．

本研究では，PCGRL における binary[2]の問題を発展させた迷路問題を扱う．binary では，初期状態の任意の 2 点の空マスの最も長い最短経路を，エージェントの行動により一定値を増加させ，固形のマスで囲まれた領域が 1 つになるように，すべての空のマスを接続するという問題である．このような問題に PCGRL を用いることで，コンテンツ生成に強化学習が有効な手法であることが示されている．また，乱数で与えられた初期状態を学習済みモデルによって，限りないパターンのコンテンツ生成が可能となることが示されている．

3. 提案手法

経路形状制御は，最短経路の形状を渦巻型になるように強化学習で制御する．渦巻型にする理由は，視覚的にも評価しやすく，強化学習によるコンテンツ生成が柔軟で強力な手法であるか調査するために扱いやすいからである．経路の情報を直進性と定義する指標で評価し，直進性を最大化することで，経路が渦巻型になるようにエージェントを学習する．

3-1. 直進性，総直進性

直進性とは，経路の情報から，最も直進した長さを示す指標である．加えて，一つ前に進んだ方向と異なる方向に進むことができる回数を設ける．これを許容回数として定義する．これらは，最短経路の制御であるから，直進性の最大化により，渦巻型の迷路の指標になる．総直進性は，許容回数を 0 から設定値までの，すべての直進性の総和である．各許容回数の直進性は，迷路が渦巻型であるか断片的に評価し，総直進性は，それらを足し合わせて総合的に評価する．

Path shape control of maze -Based on PCG via reinforcement learning-

†1 TAKAHIKO HOSHINO, Rikkyo University.
†2 YOUICHIRO MIYAKE, Rikkyo University.

3-2. アルゴリズム

図 2 に 3×3 マスの迷路の例を示す.

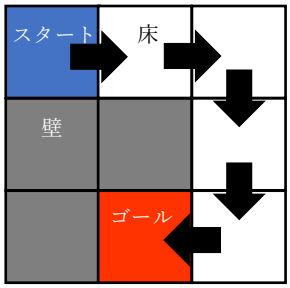

図 2 3×3 迷路の例

この例では，まず経路の情報を，

$$[[1,0],[1,0],[0,1],[0,1],[-1,0]]$$

として扱う．このように，最短経路の移動方法を取得する．次に連続で同じ方向に進んだ回数を計算する．

$$[2,2,1] \quad (1)$$

そして，許容回数が 0 であれば，直進性を単純に (1) の最大値とする．そのため，この例の直進性は 2 である．許容回数が 1 であれば，1 回だけ異なる方向に進むことができるため，(1) の隣り合う 2 つの数値を足すことで，

$$[4,3] \quad (2)$$

と情報を取得する．そして，同様に許容回数が 1 の場合の直進性は (2) の最大値である 4 と導ける．許容回数が 2 の場合，(1) の隣り合う 3 つの数値を足すことで，直進性は 5 となる．許容回数が 0 であれば，最も長い線分の長さ，1 であれば，最も長い角をなす線分の長さ，3 であれば，ゴールを囲む線分の長さのように断片的に評価し，直進性の総和により経路の形状を総合的に評価する．

4. 実験

実験設定

強化学習は PPO を用いる．学習時の変更率は 0.5 として，表現モジュールは narrow を使用し，すべてのマスを順番に移動する．エージェントの行動は，現在の位置のマスの壁マスか床マスを変更するか，スキップである．報酬は，壁マスで囲まれた領域を 1 に近づけた分の値の 2 倍と，総直進性の値と，直進性が最大値に到達した際にその直進性の 100 倍の値である．5×5 の迷路の学習ステップ数は 100 万ステップ，最大許容回数は 4 回，9×9 の迷路では 500 万ステップ，最大許容回数は 8 回である．なお，迷路の最も左上のマスをスタートマス，中央のマスをゴールマスとして扱う．図 3 にエージェントが 5×5 の迷路の修正をしている様子を示す．黄色のマスは床マスかつ最短経路のマスである．赤枠は現在の位置である．

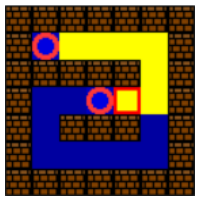

図 3 5×5 の迷路を修正する様子

実験結果

5×5 の迷路では，学習時に 15010 エピソードに対して，1028 エピソードで直進性の最大値 16 に到達した．9×9 の迷路では，38812 エピソードで一度も直進性の最大値 48 に到達できず，直進性は 33 が最高値であった．図 4 と図 5 に得られた報酬のグラフを示す．縦軸は報酬値であり，横軸はステップ数であり，Smoothing は 0.999 である．

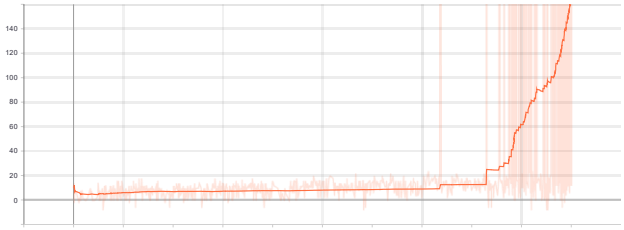

図 4 5×5 迷路のエピソードごとの報酬

図 5 9×9 迷路のエピソードごとの報酬

5. 考察

本報告では，経路形状制御問題を提案し，PCGRL を用いて実験を行った．5×5 の迷路では，今回提案した手法でも上手く学習が進むことが確認できたが，9×9 の迷路では，報酬値に増減が見られ，直進性の最大値に一度も到達できなかった．状態の次元が大きい場合，最大の直進性に到達する難易度も高く，良い方策を見つけることが困難だと考えられる．迷路の難易度を上げた際に，最大の直進性に到達できなくとも，目標に対して中間的に評価する仕組み等が，渦巻型の迷路の経路形状制御問題の今後の課題として挙げられる．

参考文献

[1] Sebastian Risi, Julian Togelius, Increasing Generality in Machine Learning through Procedural Content Generation, arXiv:1911.13071v2, 2020

[2] Ahmed Khalifa, Philip Bontrager, Sam Earle, Julian Togelius, PCGRL: Procedural Content Generation via Reinforcement Learning, arXiv:2001.09212v3, 2020