非均質マルチエージェント強化学習による障害物回避

函館高専　○高瀬　暁央，竹原　直美，石若　裕子，　北大工　嘉数　侑昇

要　旨

本研究の目的は、強化学習を用いて各エージェントが互いに協調し、自律的に目的地までの経路を学習することである。本研究ではピアノ問題を対象とし、木構造によって経路選択をするトレーニングエージェントと、Q−Learningを用いて衝突回避を行うラーニングエージェントの階層型のマルチエージェントシステムを提案する。2次元連続空間においてシミュレーション実験を行い、結果としてエージェントの軌跡と経路を示す。

1. はじめに

ピアノ問題とは、物体を部屋の外に出すために物体の回転および平行移動など、物体の姿勢を扱う問題である。この問題はSchwartzとSharir[1]により、物体が通過する空間の形状が定まっている場合においては、数学的に最適経路の計算が可能であることが証明されている。

本研究では物体を、内部に単純な行動を取ることが出来るエージェント群をもつエージェントとしてとらえる。単体のロボット内の自律分散型マルチエージェントとして協調行動を獲得するための手法を提案する。エージェント群が環境を観測して、内部状態を互いに通信することにより、衝突を回避しピアノ問題を解くことを期待する。

2. 問題空間の記述とマルチエージェントにおける学習

ピアノ問題における空間は、2次元連続空間である。狭い通路状の空間に対し、エージェントが壁に衝突することなく、姿勢を制御して初期状態から終端状態まで移動することを目的とする。

3.マルチエージェントの構成

エージェントは、外側にある物体全体を示すトレーニングエージェント（Training Agent）と、トレーニングエージェント内にあるラーニングエージェント（Learning Agent）の2種類とする。各エージェントの関係とセンサの配置を**Fig.1**に示す。

3.1 トレーニングエージェント（Training Agent）

トレーニングエージェントは長方形とし、その頂点4箇所にラーニングエージェントを持つ。エピソード毎に経路を探索し、その経路に従って大まかな移動方向をラーニングエージェントに示す。

Fig.1 The shape of the Learning Agents and Training Agent, and allocation (The line shows the sensors of Learning Agents)

経路探索には木構造を採用し、1つのノードから4方向に枝をのばしていく。のばす方向は順番に、ゴール方向、伸ばしてきた方向と同じ方向、左右にθの角度だけ曲げた方向の4方向とする。但し、左右の順番はゴールに近い方を先に探索する。トレーニングエージェントは、通行可か通行不可の2種類の状態を持つ。枝をのばすにつれ、通行不可の状態を増やす事で経路を探索する。枝をのばした方向に障害物があった場合、障害物にぶつかる枝を通行不可とする。探索木の様子を、**Fig.2**に示す。

ゴールまでの経路が存在するにもかかわらず、通路の幅がエージェントよりも狭いために通過できなかった場合は、ラーニングエージェントが通行不可の状態を作る事で、経路を修正する。

トレーニングエージェントは、次に行くべきノードの座標方向を、ラーニングエージェントに与える。エージェントが次に行くべきノードの座標値に到達した場合、ノードの階層を次へ進める。

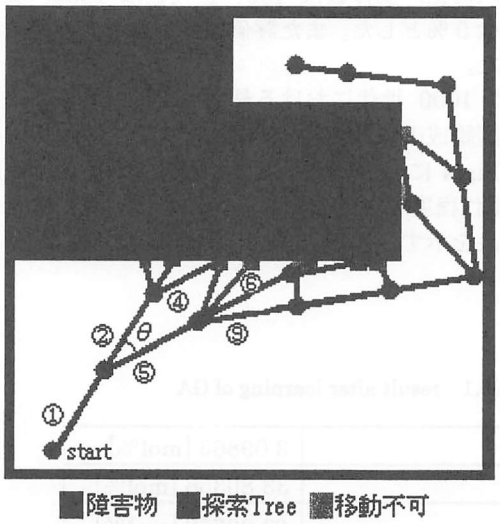


Fig.2 search by tree (goal is leaning to the left)

3.2 ラーニングエージェント（Learning Agent）

ラーニングエージェント（Learning Agent）は、センサ（Sensor）から環境を観測する。自分以外のラーニングエー

ジェント群から圧縮された各エージェントのセンサ情報を受け取り、行動を Q-Learning[2] により決定し、アクチュエータ（Actuator）に行動を出力する。他のラーニングエージェントも同様に行動を出力する。圧縮情報として、自分のセンサに１つでも障害物の反応があるかないかを、他のエージェントに通信する。ラーニングエージェントは状態として５方向のセンサ{s0,s1,…,s4}と圧縮情報を持つ。センサ範囲内では障害物に対して６段階の距離を認識することができ、各エージェントが選択可能な行動は４方向とする。
ラーニングエージェント群における Q 値の更新式は、吉田ら[3]の内部エージェントと同様に行う。

環境からの報酬は、ラーニングエージェント群が選択した行動でエージェントが障害物に衝突した、または全く動かなかった場合に、環境からペナルティを与えられる。

また、最大ステップになっても終端状態へ到達できなかった場合、近くの状態を観察し、壁の反応があるラーニングエージェントのまわりの座標を、通行不可とするようにトレーニングエージェントへ伝える。

4. シミュレーション実験

ゴールまでの通路が２つある空間で、ゴールまでの距離が近い通路はエージェントが通過できないという問題を扱い、シミュレーションを行った。

パラメータを以下に示す。問題空間の大きさを 500×500、最大ステップ数は 5000、エピソード数は 10000 とした。

トレーニングエージェントは長方形（縦 20×横 40）とし、経路探索に使うノード間の距離を 23、θ を 19、枝が障害物にぶつかった時は前後左右１ピクセルを通行不可とした。

ラーニングエージェント群はトレーニングエージェントの頂点に４個配置し、センサのレンジを 30、感度を 0~5 の６段階とした。トレーニングエージェントの質量 M を 1.0、１ステップにおける時間は 0.5、各エージェントの移動量は４とした。また Q 値については、**Fig.3** に示す環境において、Q-Learning のパラメータを学習率 α=0.1、割引率 γ=0.9、Q 値の初期値=2.0 とし、報酬は、衝突時に−1/5000、停止時に−1/5000 としてローテーションで学習した Q 値を使用した。最大ステップ数に達した時、各ラーニングエージェントはセンサを観測し、反応無しを５とした時の２以下だった時、エージェントの中心から前後左右２ピクセルを通行不可とする。

トレーニングエージェントが探索した経路の推移を **Fig.4** に示す。**Fig.5(a)**に選択した経路と実際のエージェントの軌跡を、**Fig5(b)**はトレーニングエージェントの機構に TD 法を用いた場合の軌跡を示す。

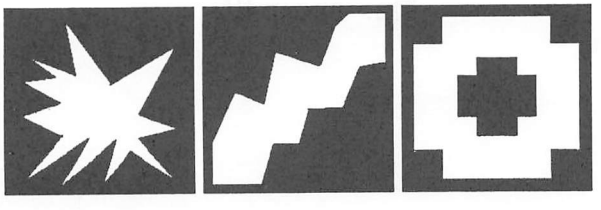


Fig.3 Environment setting Learning agents

Fig.4、**Fig.5** において、左の経路はエージェントが通れないために、経路選択を行わなくなった。エージェントの軌跡も経路選択とほぼ重なり、TD 法での経路選択に比べて、ゴールへ真っ直ぐ向かう軌跡になった。

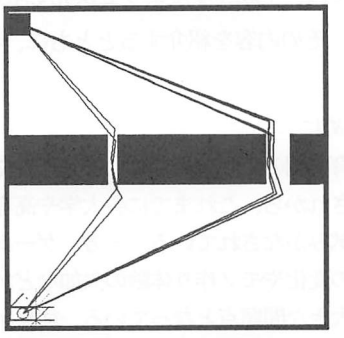
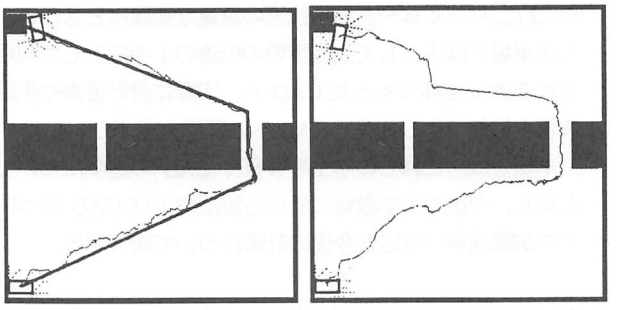


Fig.4 transition of a path found by training agent



(a) Searching tree inside of training agent　(b) TD method inside of training agent

Fig.5 Trajectories of the center coordinates of Learning agent

5. おわりに

本研究では、強化学習を用いた自律分散型エージェントにおけるピアノ問題の協調行動獲得の手法を提案し、シミュレーションにより本手法のピアノ問題へのアプローチに対する有効性を示した。今後の課題として、トレーニングエージェントにおける経路探索方法の改善を検討していきたい。

参考文献

1) Jacob T. Schwartz and Micha Sharir：“On the Piano Movers' Problem: II .General Techniques for Computing Topological Properties of Real Algebraic Manifolds, Planning, Geometry, and Complexity of Robot Motion”, pp. 51−96, 1983

2) Watkins,C.J.H. and Dayan,P.: Technical note: Q-learning, Machine Learning, Vol.8,pp.55-68,1992

3) 吉田智博,石若裕子,横井浩史,嘉数侑昇：ピアノ問題における強化学習を用いた自律分散型エージェントの協調行動獲得