マルチエージェント強化学習に基づく エレベータ制御のための報酬分配機構について

新野賢央　保知良暢　大囿忠親　新谷虎松
名古屋工業大学　知能情報システム学科
e-mail: {kenou, bochi, ozono, tora}@ics.nitech.ac.jp

1　はじめに

本研究では，マルチエージェント強化学習の枠組みにより，エレベータ制御問題に取り組む．エレベータ制御問題には，評価の方法として，たとえば，単位時間当たりの運搬量の最大化や最大待ち時間の最小化など，複数の項目が挙げられるという特徴がある．このようなドメインにおいて，すべての項目を満たすような学習をさせるためのゴール状態を，一つ定義することは困難である．また，何か一つのゴール状態の定義により，良い結果を出すような政策を獲得することは，必然的に他の評価値を下げるということが予想される．本研究では，このようなゴール状態の定義困難なドメインにおいて，トップダウン的な協調行動を取り入れた手法を提案し，問題の解決を狙う．

2　提案手法

本研究では，エージェントの振る舞いを学習過程と政策実行過程に明示的に分ける．学習過程でのエージェントの行動は，それ以前に得た経験をまったく反映せず，すべてランダムな行動をとり，Profit Sharing による学習を行う．政策実行過程では，学習過程において獲得した政策をマックス選択より実行する．

提案手法の枠組みについて述べる．まず，学習過程において，エージェントは，協調行動のサブタスク的な政策の学習を行う．そして，政策実行過程において，協調アルゴリズムに基づき個々のエージェントが内部的な政策変換を行うことで，マルチエージェントの協調行動を実現する．

2.1　学習アルゴリズム

学習過程において，各エージェントは複数の学習器を内部に持ち，並列的に学習を行う．それぞれの学習は，いずれも協調行動のいったんとなるべき政策の学習であり，独立したゴール状態をもち，独立した政策へと収束する．本枠組みでは，それら複数の評価，一つ一つに長けた政策を同時並列的に学習する．

Reward Sharing Mechanism for Elevator Control Base on Multi-Agent Reinforcement Learning
Katsuhisa SHINNO, Yoshinobu BOCHI, Tadachika OZONO, Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya 466-8555 JAPAN

実行過程においてエージェントは，協調アルゴリズムに基づき内部的な政策変換を行う．これにより，協調的なサブタスクをエージェントごとに実行し，全体としての協調行動を実現する．

2.2　協調アルゴリズム

本提案手法では，学習過程において得た複数の政策を組み合わせることで，協調行動の実現を目指す．本手法における協調アルゴリズムとは，エージェントがどの政策をどのタイミングで行うのかを決定するものである．本研究では以下の２通りのアルゴリズムをもって，その有用性を検証していく．

確率的な政策変換：

個々のエージェントが設定された確率に基づき，独立して政策を変更する．

状況判断による政策変換：

政策変換の必要な状況を外部の情報から判定し，ルールに見合う状況において政策を変更する．

3　実験

本研究では，提案手法をエレベータ制御に適用する．

エレベータの稼動する環境は，エレベータの数は２基，７階建てのビルを想定する．シミュレーションは論理時間を取り入れた連続時間で行う．したがって，停止，発進，乗降作業などの各挙動にはそれぞれ時間的コストを伴う．

ビルのフロアからは論理単位時間当たりに一人の乗客が発生する確率を設定する．各フロアの確率は，１階が 10%，３階が１%，５階が１%，７階が４%とし，他のフロアは非常に小さな値を設定する．

強化学習エージェントはエレベータとし，エージェントの取れる行動を上下の加速動作と何もしないという３つを定義する．エージェントは自分と他のエージェントの速度，位置と，エレベータの要求順序をあらわす変数を知覚する．

提案手法におけるサブタスク的な学習は,「平均待ち時間の最小化」を狙う学習と「最大待ち時間の最小化」を狙う学習とする．前者の報酬発生条件は，要求のあ

るフロアに停止したときで，報酬量は乗せた人に比例する量を与える．後者では，要求順序の一番古いフロアに停止したときのみ，一定量の報酬を与える．

政策実行過程において，２基のデフォルトは「平均待ち時間の最小化」を対象とする政策をとる．確率的な政策変換方式においては，２基のうち１基を政策実行タームの 100 論理時間ごとに 50%の確率で政策を変更する．

状況判断による政策変換方式では，乗客のうち最大待ち時間が 130 論理時間の閾値を超えると 1 基のエレベータが政策を変更する．変更している時間は閾値を超える乗客がいなくなるまでとする．

このような設定において，実験では学習過程と政策実行過程を 1000 論理時間ごとに繰り返して行った．実験結果を以下の図 1，図 2 に示す．X 軸のステップ数とは，繰り返した学習過程の数をあらわす．Y 軸の時間の単位は「論理時間」である．図 1 の平均待ち時間とは運搬した乗客の総待ち時間を運搬した乗客数で割った値である．また，図 2 の最大待ち時間は，その政策の実行中に最も待たされた人の待ち時間である．

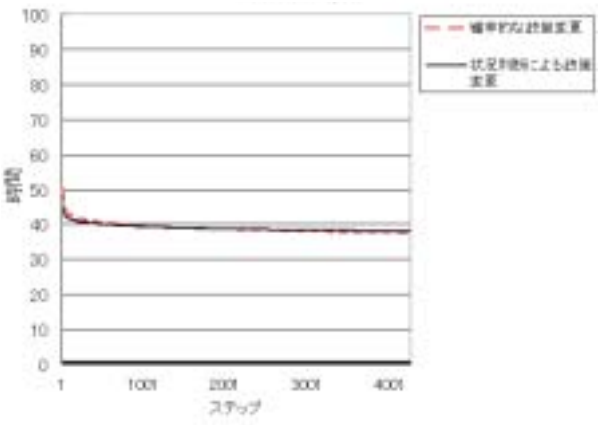


図 1: 平均待ち時間の変化

4　評価・考察

図 1 の平均待ち時間のグラフをみると，二つのアルゴリズムは同じ値に収束している．これは両アルゴリズムとも，乗客の発生に対して十分な運搬性能を発揮していると考えられる．

図 2 の最大待ち時間のグラフでははっきりした違いが見られた．状況判断による政策変更は，政策変更を閾値により決定知るため，最悪に待たされる状況を確実に回避することができる．そのため，閾値の 130 論

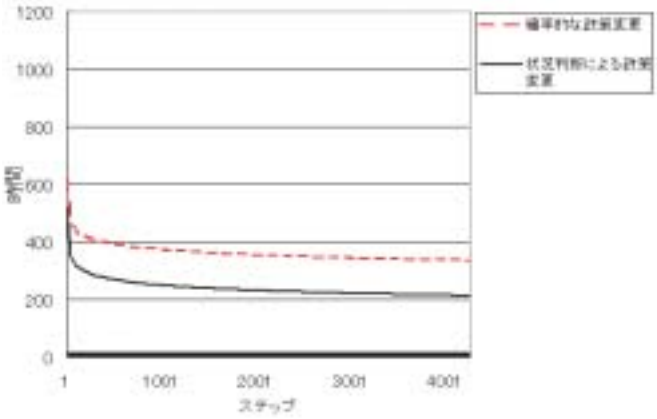


図 2: 最大待ち時間の変化

理時間に対し，高々200 論理時間という値に収束している．これに対し，確率的な政策変更は，必要なときに政策の変更がされない負事例の影響により，他方より劣る結果となったと考えられる．

5　まとめ

複雑な現実問題に対して強化学習による解決を試みるとき，評価方法が複数ある場合など，ゴール状態が安易に定義しがたい場合が多い．

本研究では，問題を複数のサブタスクの組み合わせとして定義する．学習過程においてエージェントはこれらの学習を並列的に行い，内部で複数の政策を獲得させた．エージェントは政策実行過程において，協調アルゴリズムに基づき，内部的に政策を変換する．このようなトップダウン的に協調行動を実現することで，本来一つのゴール状態を定義できないドメインに対して，より有効な政策を獲得できるという手法を提案した．

本研究のように，ボトムアップ的な学習手法である強化学習にトップダウン的な枠組みを導入することは，現実世界に応用する上で重要な試みといえる．本研究では，エレベータ制御問題において，提案手法を適用した．シミュレータ上での実験を通して，提案手法のアルゴリズムを考察した上で，その有用性を確かめることができた．

参考文献

[1] 荒井幸代，宮崎和光，小林重信 “マルチエージェント強化学習の方法論〜Q-learning と Profit Sharing による接近,” 人工知能学会誌，Vol.13，No.5，pp.609–618，1998.