クラスター分析を用いたエージェントの状態空間構成法

有福直也†，梶川嘉延†,††，野村康雄†
関西大学工学部電子工学科†，関西大学学術フロンティアセンター††

1. はじめに

近年，環境とエージェントの相互作用を通して学習する手法として強化学習や記憶に基づくロボット学習が注目されている．しかし，実世界性を持ったシステムにおけるエージェントタスクに強化学習を適用する場合，最も困難な問題の一つは状態空間の構成である．近年では，ゴール状態に一つの行動で到達できる状態を集合化し,順にゴールの一つ手前の状態に到達できる行動の状態を集合化する．これをスタート位置まで求めることにより最適な状態を作成する方法[1]や，局所予測モデルを用いてエージェント間に及ぼす影響を推定しエージェントの分類を行い，その後強化学習によりタスク達成率を検討している手法[2]等が提案されている．

一方，本稿では各状態空間のデータに基づき，クラスター分析を用いて状態空間の構成を行う．これにより，状態をあらかじめ分割しておいても状態の「類似度」により状態を統合でき，環境に適した状態の構成を行うことができる．

2. 環境設定

本稿では，**Fig.1** に示す格子面を設定し，エージェントがゴールに到着する問題から提案法の有効性を検討する．これは，格子問題では連続状態を離散化する場合よりも比較的容易に扱うことができるからである．なお，エージェントの行動には強化学習法の Profit Sharing を用いる．

3. クラスター分析を用いた状態空間構成

3.1 エージェントの状態と基本行動

まず，エージェントの基本行動について説明する．エージェントは **2** で説明した格子面を上下左右の 4 方向に移動する．なお，**Fig.1** において(i,j)=(8,8)，ゴール位置はランダム，エージェントの初期位置は格子面の一番外側，行動選択にはルールの重みによるルーレット選択を用いる．

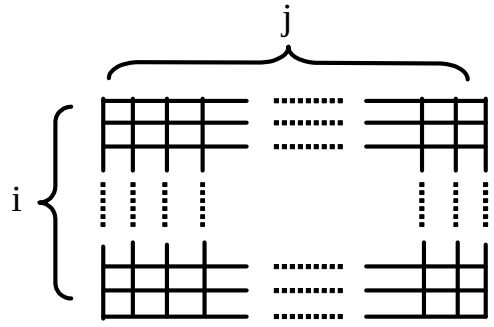

Fig. 1 Simulation Environment.

3.2 クラスター分析に用いるデータの設定

次に，クラスター分析を用いた状態空間の構成について説明する．これは初期の状態分割から状態に対する「類似度」を求め，同じような状態を結合させる方法である．始めに，使用する対象データ $D_n\{n=0,\ldots,63\}$ を以下のように設定する．

$$D_n=\{\boldsymbol{S_n}, \boldsymbol{V_n}, R_n\}$$
$$\mathbf{S_n}=[s_{nu}, s_{nd}, s_{nl}, s_{nr}]$$
$$\mathbf{V_n}=[V_{d,n}, V_{\theta,n}]$$

ここで，$\boldsymbol{S_n}$ は各状態の隣接状態数を表し，状態が四方に隣接している状態を保存する．$\boldsymbol{V_n}$ は各状態に対する移動方向を単位円上のベクトルに変換し，それを正規化した大きさ $V_{d,n}$ の和とその角度 $V_{\theta,n}$ をまとめたものである．R_n は報酬更新回数を表し，各状態で正の報酬を得ることができれば+1 を，負の報酬の場合は-1 を加える．

以上を状態のデータとし，これを用いて状態空間を構成する．

3.3 クラスター分析の使用方法

クラスター分析の手順は

1. 状態 s_n で隣接状態 $\boldsymbol{S_n}$ に対してクラスター分析を行い，結合候補 s_n' を得る．
2. s_n' の隣接状態に対してクラスター分析を行い，結合候補 s_n'' を導出．$s_n = s_n''$ であれば，s_n と s_n'' の状態を結合し終了．そうでなければ 3 に進む．

「A Study on Agent's State Space Generation Using the Cluster Analysis」

† 「Naoya Arifuku, Yoshinobu Kajikawa, Yasuo Nomura • Department of Electronic, Faculty of Engineering, Kansai University」
†† 「Yoshinobu Kajikawa • Frontier Science Center, Kansai University」

3. s_n''に対して 2 と同様の処理を行い，s_n''の状態候補s_n'''を得る．このとき，$s_n = s_n'''$であれば 4 へ進む．また，$s_n' = s_n'''$であればs_n'とs_n'''を結合し終了．そうでなければ，s_n'にs_n''を，s_n''にs_n'''を代入し，3 の処理を繰り返す．
4. 各状態候補間の相関係数rを求め，$r > 0.9$の結合候補を抽出する．そして，その中で最もrが大きい候補を結合する．そうでなければ，結合しない．

これによりクラスターの中で最も「類似度」の大きい状態を結合することができる．なお，結合後にできた状態のD_{new}は統合した 2 つの状態のD_nを平均したものとする．

4. 状態再構成の使用時期

提案法により状態の再構成を行うタイミングは，学習を行っている最中に行動の中で得られる情報だけでそれを決定する方法が良いと考えられる．そこで，文献[3]で用いられている“学習残エントロピー”を利用する．これは式(1)で表される量である．

各状態s_nに対して，行動$a_k\{k = 0,\ldots,3\}$に対する行動選択確率$p(a_k, s_n)$から

$$I(s_n) = -(1/\log 4)\sum_a p(a_k, s_n)\log p(a_k, s_n) \quad (1)$$

を定義する．これは，各状態において行動の不確定性（自由度）がどの程度残っているかという指標である．

また，エピソードごとにエピソードEに含まれている状態全てについて$I(s_n)$を平均した学習残エントロピーIを式(2)で定義する．

$$I = \frac{1}{|E|}\sum_{s_n \in E} I(s_n) \quad (2)$$

ただし，$|E|$はエピソードEに含まれる状態数である．提案法による状態結合を行う場合，学習が適度に行われた状態で使用することが良いと考えられる．そこで，良好な結果であった$I \le 0.6$の場合に状態空間構成を行うこととする．

5. シミュレーション結果

Fig.2 にシミュレーション結果を示す．この結果から，提案手法を用いた場合が従来法に比べ学習の収束が向上していることが分かる．これは，提案法による状態再構成により状態数が削減し，不適切な行動を抑制できたためと考えられる．また，**Fig.3** の学習残エントロピー値に関しても提案法は従来法よりも減少が大きく，状態の結合により行動の不確定性が少なくなったと考えられる．

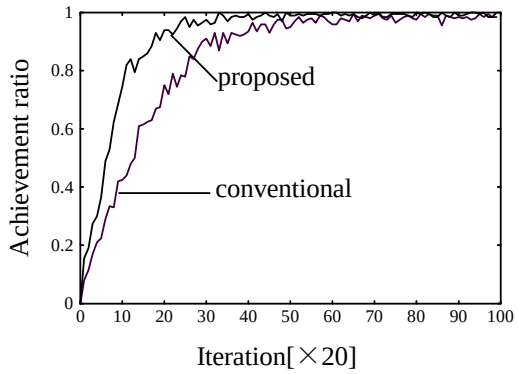

Fig. 2 Simulation result (Achievement ratio).

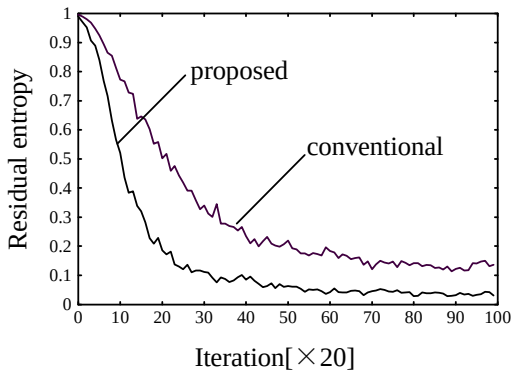

Fig. 3 Simulation result (Residual entropy).

6. まとめ

事前に離散化した状態から，学習を通じて状態を再構成する手法を提案し，その結果学習速度が向上することを示した．

今回は簡単な環境でシミュレーションを行ったが，今後は環境をより複雑化した場合の提案法の有効性，学習残エントロピー値の使用方法について検討の余地がある．

参考文献

[1]浅田 稔, 野田 彰一, 細田 耕, “ロボットの行動獲得のための状態空間の自律構成”, 日本ロボット学会誌, Vol. 15, No. 6, pp. 886-892, Jun. 1997.
[2]内田 英治, 浅田 稔, 細田 耕, “マルチエージェント環境における部分区間同定法を用いた状態空間の構成と行動獲得”, 第 15 回日本ロボット学会学術講演会予稿集, pp. 895-896, 1997.
[3]伊藤 昭, 金渕 満, “知覚制限の粗視化によるマルチエージェント強化学習の高速化”, 信学論 D-1, Vol. J84-D-I, No. 3, pp.285-293, Mar. 2001.