動的な階層環境における強化学習エージェントの確率知識を用いた方策改善に関する研究

動的な階層環境における強化学習エージェントの確率的知識を用いた方策改善に関する研究

室蘭工業大学 大学院工学研究科 博士後期課程
生産情報システム工学専攻

UTHAI　PHOMMASAK

目 次

第1章　序論

1.1　研究背景

最近、救助ロボットは、地震や津波などの大災害のときに犠牲者を救い、かつ災害後のクリーンアップ作業に人間の危険を減らすために必要となっている．したがって，様々な環境を適応できるようにロボット工学を発展させることもさらに必要になる．また，コンピュータの高度化に伴い，宇宙開発，さらにエンターテインメント等への実用化に向けた研究が盛んになっており，実質的な業務を果たすロボットも登場している．その中で，ロボットが自ら環境の情報を獲得し，動作を計画，実行する自律制御システムへの需要が高まっており，ノイズや環境変化等の不確実性が存在する実環境における，ロボットの適応的学習の実現に向けた研究が行われている．これを背景として，学習主体者が環境と相互作用し，情報の獲得と行動を選択する学習法が改めて注目されており，その中でも機械学習の一つである強化学習は，簡潔なアルゴリズムと強化な数字的基礎に支えられ，さまざま応用が期待されている手法である [19, 20].

強化学習においては，エージェントには学習のための正しい行動を教示されず，環境との相互作用を通じて自身の方策を最適化していくことで，対象となる環境に適応する．相互作用によってエージェントが観測するデータは，環境に関する情報を表していると考えられ，これを教師信号的な役割として方策の改善に適用する研究が行われている．

強化学習は，文字通り環境を広く探索して環境についてのモデルを構築することによって適切な方策の学習を行う環境同定型と，環境の探索をある程度犠牲にして学習途中においてもなるべく報酬を得続け，報酬を得た経験を強化するという経験強化型の２つを大きく分けられ

る．環境同定型の代表であるQ-学習などは，報酬の伝搬が一段階ずつしか行なわれないことにより学習速度が遅い，また，環境を探索するために多大な計算時間がかかるため，本研究では，報酬を利用したデータ系列の重み値を一括更新により方策の学習を行う学習速度が速い経験強化型型の利益共有法を採用する．

通常，強化学習エージェントは新しい環境におかれると，方策を初期化して一から学習をやり直す必要があるため，環境変化の認識と，環境変化後の適応のための適切な方策改善処理を実施し，未知環境を含む環境変化に効率的に適応可能となるような手法はたくさん開発されている [1, 2]．ただし，動的な環境や3次元の環境などの複雑な環境に適応するときに，入力状態や出力行動の数を含む実験パラメータの設定によって，エージェントが方策をうまく学習できない場合もある．また，このように強化学習の枠組みの外部から方策改善を適用することの有効性が存在する一方で，システムの複雑化に伴い計算量は増大し，計算資源に制限のある実環境を想定した場合には計算量の抑制は重要な問題となる．

1.2 研究目的

強化学習エージェントが過去に学習した環境の観測データ，環境に関する情報，すなわち知識であるといえる．エージェントが自ら獲得した複数の既知環境に対応する知識を混合し，教師信号として適用することで，未知環境へ効率的に適応可能となることが期待される．

強化学習エージェントの入力状態を決定する観測方向と出力行動の方向を増やしながらも，利益共有法を基礎にして、複雑な実環境に適応できるように報酬与え方や重みの更新式などの改良を加えた手法を提案する [11, 12]．その中では、階層型環境適応用のエピソードの二次元化するなどのパラメータの新たな更新法の導入を行う．加えて、エージェントの観測データからなる同時分布を構成要素として混合分布 [13] を用いて，未知環境に適応できるように方策改善を行うための

手法を構築する．さらに，方策改善性能を維持しながらも計算量を抑制するために，混合分布の構成要素を統計的手法に基づいて少ない数で選択できる分布クラスタリングを新たに導入する．本研究においては、計算量を抑えながら方策改善性能を保ち、動的な環境階層型環境に適応可能な効率的な強化学習システムの提案を通じて、未知環境への対応と実用性を備えた強化学習手法を構築し、ロボット制御におけるアルゴリズム分野に貢献することを目的としている

1.3　本論文の構成

まず第 2 章で本研究の基盤となる強化学習とその一手法である利益共有法（Profit sharing），そして，利益共有法のパラメータの更新について説明する．次に第 3 章に本研究で提案する混合分布を用いた方策改善について述べ，第 4 章ではクラスタリングによる混合分布の構成要素の選択について説明する．第 5 章に本研究（システム）の流れについて説明する．第 6 章で提案手法について計算機実験と考察を行い，最後に第 7 章でまとめとする．

第2章　強化学習

2.1　強化学習とは

強化学習 (Reinforcement Learning) とは，ある環境内におけるエージェントが，現在の状態を観測し，取るべき行動を決定する問題を扱う機械学習の一種．図 2.1 のようにエージェントは行動を選択することで環境から報酬を得る．強化学習は一連の行動を通じて報酬が最も多く得られるような方策 (policy) を学習する [7, 8, 9].

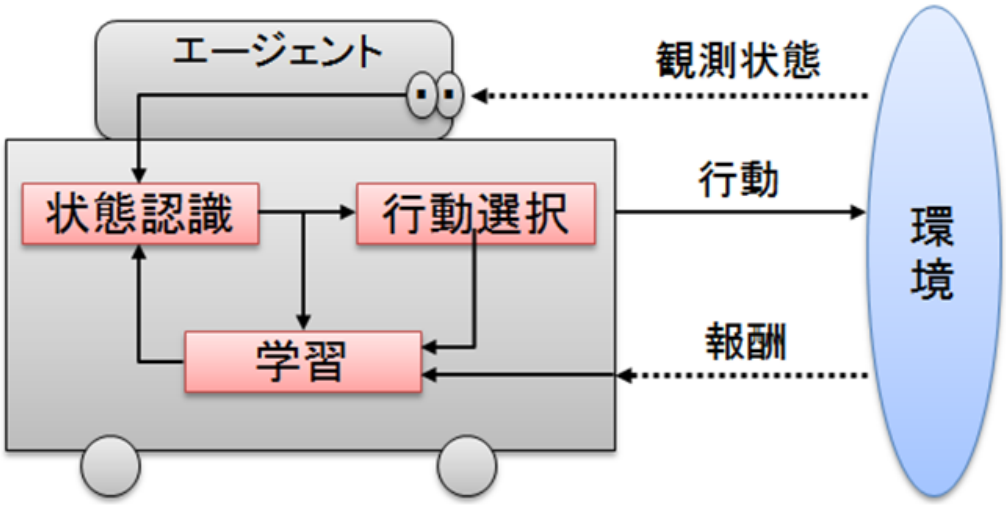


図 2.1: 強化学習のイメージ

強化学習は，学習のための適切な入力データと出力データのペアが与えられることがない，という意味からすると，教師あり学習とは異なる学習手法である．また，未知の学習領域を開拓していく行動と，既

知の学習領域を利用していく行動とをバランス良く選択することができるという特徴も持っている．その性質から未知の環境下でのロボットの行動獲得に良く用いられる．

2.1.1 強化学習の特徴

強化学習の主な特徴としては以下のように挙げられる．

- 学習に際して教師信号を用いない．
- 行動と評価のサイクルを振り返しながら学習が進む．
- 確率的な行動規則の獲得が可能なため，ノイズの多い実環境にも対応可能である．

また，状態，行動，報酬といった単純な形式でアルゴリズムが定式化されているため，数学的な取り扱いが容易である．

2.1.2 強化学習の枠組み

学習主体者をエージェントと呼ぶ．強化学習エージェントは環境との相互作用の振り返しを通じて自身の方策 (policy) を最適化していくことで対象となる環境に適応する．環境との相互作用について具体的に述べると，

1. 環境から状態 s を観測する．
2. s に基づいて行動 a を起こす．
3. その結果として状態 s' に遷移するとともに，報酬 r を得る．

ということであり，このサイクルの振り返しにより学習が進む．得られる報酬は $r \leq 0$，すなわちペナルティに相当する場合もある．

一般に，エージェントがより多くの報酬を得るためには，環境を広く探査 (同定) しなければならない．また，これまで得た知識を用いて貪欲（greedy）に報酬の最大化を追求することも必要である [10]．し

かし，環境の同定を重視して行動すると，学習途中での報酬が軽視されがちになる．このように強化学習では，報酬獲得と環境同定といった相反する目的が要求される．強化学習の手法では，学習途中での報酬獲得を重視する手法である利益共有法等の経験強化型と，最適な方策を得るために環境同定を重視する手法である Q-learning 等の環境同定型という 2 つのアプローチに大別される．前者は主にマルコフ決定過程 (MDP) の環境への適応，後者は非マルコフ決定過程の環境への適応に用いられる．

2.2 利益共有法

本研究では，エージェントの行動決定に関して，強化学習の一手法である利益共有法を用いる．利益共有法では，報酬を利用したデータ系列の重み値の一括更新により方策の学習を行う [11, 12].

2.2.1 方策学習の方法

強化学習エージェントは観測状態 s と，その状態で出力する行動 a の対からなるルール (s, a) の重み $w(s, a)(\forall s \in S, \forall a \in A)$ の値 (≥ 0) をもとに行動を選択する．ルール (s, a) が選ばれる確率は，

$$P(rule = (s, a)) = \frac{w(s, a)}{\sum_{s' \in S, a' \in A} w(s, a')} \tag{2.1}$$

S, A はそれぞれ状態と行動の集合に表す．エージェントが選択した初期ルール (もしくは報酬獲得時に選択したルール) から次に報酬が得られるまでに選択したルール系列 $\mathrm{L} = \{(s_1, a_1), (s_2, a_2), \ldots, (s_L, a_L)\}(L$: 系列長) をエピソードと呼ぶ．

利益共有法では，エピソード中のルール重みを一括更新することで学習を行う．ルール (s_L, a_L) を選択した結果，報酬 r が得られたとすると，エピソード中の各ルール (s_i, a_i) に対する重みは以下に従って更新される．

$$w(s_i, a_i) \leftarrow w(s_i, a_i) + f(i) \tag{2.2}$$

$$f(i) = r\gamma^{L-i} \quad (i = 1, 2, \ldots, L) \quad (0 < \gamma < 1) \tag{2.3}$$

$f(i)$ はルール系列に報酬を分配する強化関数と呼ばれ，γ は学習率となる．

2.3　パラメータの更新

通常では，強化学習エージェントが上下左右の 4 方向の状態を監視し，上下左右の 4 方向の行動を出力するが，複雑な環境の場合に対応できない可能性がある. そのため，監視する方向と出力行動の方向を $(\uparrow, \downarrow, \leftarrow, \rightarrow, \nwarrow, \nearrow, \swarrow, \searrow)$ の 8 方向を増やすことで複雑な環境にも対応で可能となることが考えられる．ただし，監視する方向と出力行動の方向を増やすことで，ルールの数が $(2^4 \times 4)128$ 個から $(2^8 \times 8)2048$ 個に増大し，エージェントが目的地に到達するまで選択したルールも増大してしまう. エピソード中のルールの数が多いとエージェント最初に選択したルールの重みは更新されていないので，方策をうまく学習できない可能性がある. エージェント最初に選択したルールの重みが更新されるように，エピソード中のルールの数を減少する必要がある. エピソード中のルールの数を減少するには，エピソードの取り方を変えることと，エージェントが少ない行動選択回数で目的地に到達できるようにすることと考えられる [23].

エピソードの取り方の更新とエージェントが少ない行動選択回数で目的地に到達できるように報酬与え方の更新は以下となる.

エピソード取り方の更新:

図 2.2 のように，エージェントが目的地に到達するまで 2 回以上選択されたルールを削除し，最後に選択されたルールを 1 個だけを残す.

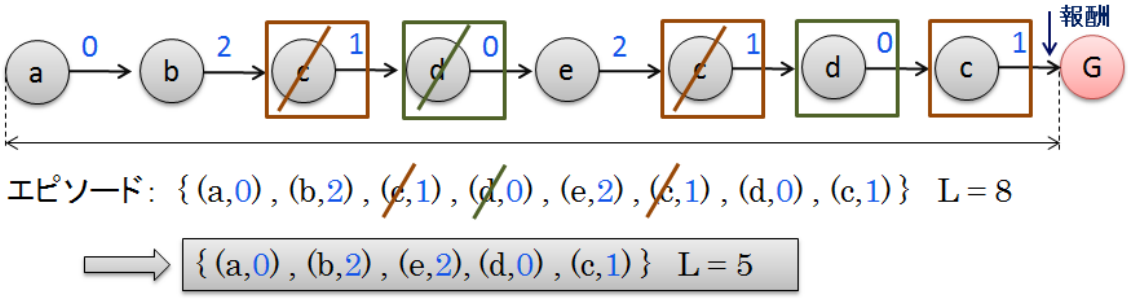


図 2.2: エピソード取り方の更新

報酬与え方の更新:

報酬の値を 固定された値から，エージェントが目的地に到達するまで行動の選択した数数によって報酬の値を決定する非固定値に更新する．目的地に到達するまでの行動選択回数が少なければ少ないほどいいということで，次式 (2.3) のように行動選択回数が少なければ少ないほどエージェントに報酬を多く与える．

$$w(s_i, a_i) \leftarrow w(s_i, a_i) + r\gamma^{L-i}$$
$$\Downarrow$$
$$w(s_i, a_i) \leftarrow w(s_i, a_i) + (r_0 + t - n)\gamma^{L-i} \tag{2.4}$$

r_0 は初期報酬，t は 1 試行の行動選択回数の制限，n は実際に行動を選択した回数となる．

2.3.1 三次元に拡張した場合の学習法

2.2.1 節では二次元環境における学習方法を示した．そこで，この学習法を基礎にして三次元環境に拡張した次の方法をとる．
今までは $z = 1$ の状態 (s, a)，つまりは $(1, s, a)$ であったが，行動 a によって上位状態 z も変化するものと定めると，ルール (z, s, a) が選ばれる確率は，

$$P(rule = (z, s, a)) = \frac{w(z, s, a)}{\sum_{s' \in S, a' \in A} w(z, s, a')} \tag{2.5}$$

である．エージェントが初期状態から報酬が得られるまでに選択したルール系列を

$$\mathcal{L} = \{(z_1, s_1, a_1), \ldots, (z_L, s_L, a_L)\} \ (L : \text{系列長}) \tag{2.6}$$

とする．報酬 r が得られたとき，エピソード中の各ルール (z_i, s_i, a_i) に対する重みは以下の式に従って更新する．

$$w(z_i, s_i, a_i) \ \leftarrow \ w(z_i, s_i, a_i) + f(i) \tag{2.7}$$

$f(i)$ については式 (2.3) と同様である. このように拡張を行うことで三次元環境での強化学習を実現することができる.

2.4 エピソードの二次元化による強化関数の改善

2.3.1 節で拡張した方法では効率的な学習を行うことができない. 何故ならば 3 次元の階層構造であると，一般的に 2 次元環境よりも報酬 r を獲得するまでの距離が長いためである. 報酬を獲得するまでの距離に比例して学習効率が低下する事は，強化学習に共通する問題点の 1 つである.

2.4.1 学習効率の低下

エピソードはエージェントが報酬 r を獲得するために選択したルール系列である. 報酬獲得までの距離が長いという事は，このエピソードが長くなるために式 (2.3) では初期に選択したルール系列ほど強化されないということになる. 報酬 r を得たとしても学習効果が著しく低下することが数式からも示されている.

エピソードの分割方法

エピソード $\mathcal{L}$ を z に関する移動選択が為されたルール毎に分割する.

$$
\begin{aligned}
\mathcal{L} &= \{(z_1, s_1, a_1), \ldots, (z_L, s_L, a_L)\} \\
&\downarrow \\
\mathcal{L} &= \left\{ \begin{array}{cccc}
(z_1, s_1, a_1) & (z_1, s_2, a_2) & \ldots & (z_1, s_{L_1}, a_{L_1}) \\
(z_2, s_1, a_1) & (z_1, s_2, a_2) & \ldots & (z_2, s_{L_2}, a_{L_2}) \\
\vdots & \vdots & \ddots & \vdots \\
(z_n, s_1, a_1) & (z_n, s_2, a_2) & \ldots & (z_n, s_{L_z}, a_{L_z})
\end{array} \right\} \quad (2.8) \\
(z_L, s_L, a_L) &= (z_n, s_{L_z}, a_{L_z}) \\
L &= \sum_{i=1}^{n} L_i \quad (2.9)
\end{aligned}
$$

上記の行列では $(z_1, s_{L_1}, a_{L_1}), (z_2, s_{L_2}, a_{L_2})$ が z に関する行動選択を意味している．z の移動が認められた場合は，次の行に移り選択されたルール系列を記憶していく．最後の (z_n, s_{L_z}, a_{L_z}) は報酬 r を獲得したルールである．
このような方法でエピソードを分割することができる．

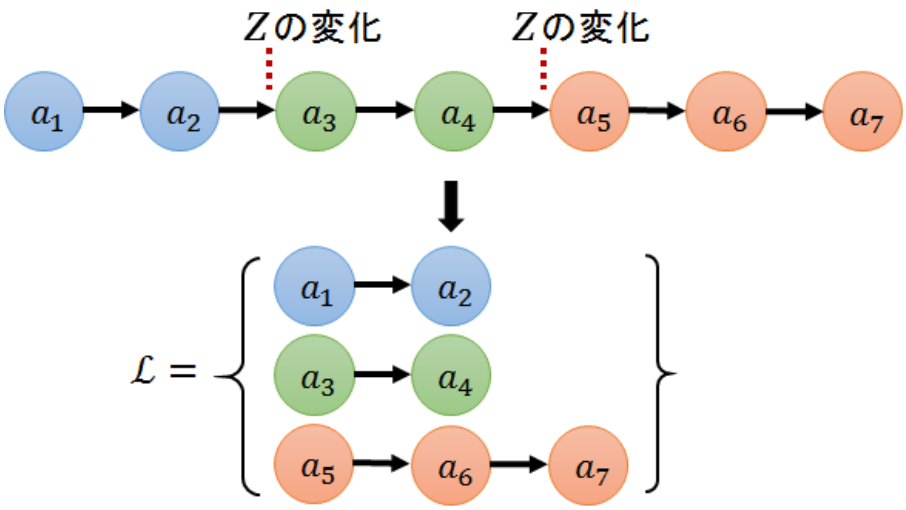


図 2.3: エピソード二次元化の例

2.4.2 強化関数の改善

エピソードを二次元化する事によって強化関数を適切に改善することが可能となる．z に関する移動により分割されることで，昇降に擬似報酬 r_i を設定することができる．

擬似報酬の設定

報酬 r を獲得したとき，分割された各エピソードには以下の式に従った報酬 r_i が与えられる．

$$r_i = r\gamma_z^{n-i} \quad (i = 1, \ldots, n) \quad , \quad (0 < \gamma_z \leq 1) \tag{2.10}$$

γ_z は報酬 r をどの程度の割合で各エピソードに渡すかを決定する上位学習率と呼ぶ.

強化関数の変更

式 (2.10) を用いて以下の式に従って重みを更新する.

$$w(z_i, s_j, a_j) \leftarrow w(z_i, s_j, a_j) + f(i, j) \tag{2.11}$$

$$\begin{aligned} f(i,j) &= r_i \gamma^{L_i - j} \\ &= r \gamma_z^{n-i} \gamma^{L_i - j} \quad (i = 1, \ldots, n), \ (j = 1, \ldots, L_i) \end{aligned} \tag{2.12}$$

式 (2.12) の通り r_i を式 (2.10) から代入して 1 つの式にまとめると，結局は r を用いて強化を行うことに変化はない．しかし，各エピソードの末尾は必ず r を獲得したルールか z の移動に関するルールである事から，本来 z の移動には報酬 r は与えられないが，擬似的な報酬 r_i が与えられていかにもそこに報酬があるかのような挙動を見せる.

擬似報酬の必要性

例えば z の移動が行える状態，即ち階段などが環境に与えられたとする．この状態に報酬 r を設定することがまず考えられる事であるが，これは避けるべきである．何故ならば，エージェントは目的を達成することで報酬 r を得るのであって，ここでの目的達成は階段を昇り降りする事ではない．階段の昇り降りは，目的達成のために必要になる可能性のある分岐点である.

例えば階段に報酬 r を与えてしまうとこのような不都合が発生する可能性がある．本来たどり着くべき目的地は，階段を昇らず通過した先にある場合，エージェントは階段を見つけたら昇る事により報酬 r を獲得してしまい，この行動が強化されることによって，目的地にたどり着けないで学習が停滞する危険性がある．よって，階段を昇るという行為が本当に必要であったかどうかを，報酬 r を用いて判定すべきなのである．このような理由から，擬似報酬 r_i を報酬 r を用いて表現することにより，上記の危険を回避している.

2.5　無効エピソード問題

2.5.1　無効エピソードとは

前節2.4.2にて強化関数の改善手法を示した．しかし，この方法を取ることによって１つの問題点が浮上する．あるルール(z_1, s_{L_1}, a_{L_1})を選択して(z_2, s_1)に移動したとする．ここでのエピソード$\mathcal{L}_2 = \{(z_2, s_1, a_1), \ldots, (z_2, s_{L_2}, a_{L_2})\}$によって，元の状態$(z_1, s_{L_1})$に戻る場合が考えられる．これらのルール系列は報酬獲得に貢献しない可能性があり，このような迂回系列全体を無効エピソードと呼ぶことにする．

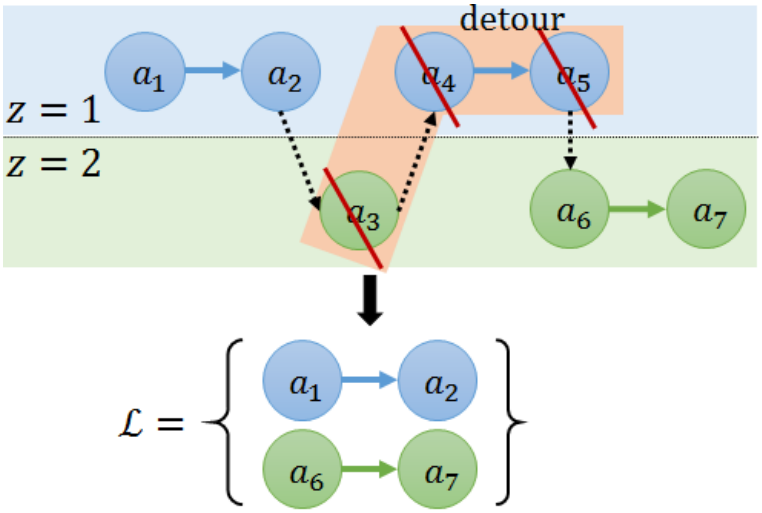


図 2.4: 無効エピソードの判別例

2.5.2　無効エピソードの抑制

前小節 2.5.1 のような無効エピソードには他の危険性も存在する．zの移動に関するルールがアルゴリズム上，比較的大きく強化されることになるが，zの移動に関するルールが繰り返し選択され続け，そこから脱出できずに学習が停滞する危険性を含んでいる．

このような事からも無効エピソードを抑制・排除しなければならない. そこで以下のような方法で無効エピソードを解決することを提案する.

あるエピソード $\mathcal{L}_i$ のエピソード長 L_i が

$$L_i \leq L_c \text{ かつ } z_{i-1} = z_{i+1}$$

であるとき，$\mathcal{L}_i$ は無効エピソードであると定める. このとき，無効エピソードの中のルール系列全てと $\mathcal{L}_{i-1}$ の末尾のルール $(z_{i-1}, s_{L_{i-1}}, a_{L_{i-1}})$ を除外する. L_c は無効エピソードを判定するための定数である.

第3章　混合分布による方策改善

3.1　混合分布とは

エージェントが方策学習中に観測するルール系列からなる同時分布 $P(z, s, a)$ は，環境に関する確率的な知識といえる．そこで，複数の既知環境で得られる $P_i(i = 1, \ldots, M)$ の混合分布を利用し，エージェントが獲得した方策の改善を行う．混合分布は次式によって表される．

$$P^{mix}(z, s, a) = \sum_{i=1}^{m} \beta_i P_i(z, s, a) \tag{3.1}$$

ここで m は同時分布の総数，β_i は混合パラメータを表す $(\sum_i \beta_i = 1, \beta_i \geq 0)$．この混合パラメータを学習の対象となる環境に応じて調節することで，未知環境における方策を適切に改善可能となることが期待される．

3.2　混合パラメータの決定方法

$P_i(z, s, a)$ は確率分布であるため，本研究では，確率分布の距離として統計的評価でよくりようされる距離関数 Hellinger distance[**?**] を用いる．

$$D_H(P_i, Q) = \left\{ \sum_x \left[P_i(x)^{\frac{1}{2}} - Q(x)^{\frac{1}{2}} \right]^2 \right\}^{\frac{1}{2}} \tag{3.2}$$

D_H は分布 P_i, Q 間の距離を表し，同一である場合は 0 となる．P_i はエージェントが過去に学習した m 個の環境で得られた同時分布であり，Q は未知環境で τ 回の試行で得られたサンプルの分布である．

$$\begin{aligned} D_H(P,Q) &= \sqrt{\sum P(x)^2 + \sum Q(x)^2 - 2\sum P(x)Q(x)} \\ &= \sqrt{1+1-2\sum P(x)Q(x)} \\ &= \sqrt{2-2\sum P(x)Q(x)} \end{aligned} \tag{3.3}$$

式 3.2 により，D_H の最大値が $\sqrt{2}$ であることから，次式によって混合パラメータを決定できる．

$$\beta_i = \frac{\sqrt{2} - D_H(P_i, Q)}{\sum_{j=1}^{M}(\sqrt{2} - D_{H_j})} \ (i = 1, \ldots, M) \tag{3.4}$$

但し $\sum_{j=1}^{M}(\sqrt{2} - D_{H_j}) = 0$ のとき，$\beta_i = \frac{1}{M}$ とする．すなわち，全ての分布が同一であった場合，混合パラメータは平均に割り振られる．また，D_H は距離の公理を満たすことにより [4]，第 4 章で説明するクラスタリングの距離関数としても利用する．

3.3 方策改善の手順

混合分布の構成と方策改善は以下の手順によって行われる．

1. エージェントが過去に利益共有法により学習した M 個の環境に対応する同時分布 $P_i(i = 1, \ldots, M)$ と，現在の環境（未知環境）で τ 回の方策学習の間に得られたサンプルの同時分布 Q の D_H を求める．
2. 混合パラメータ β_i を求める．

3. 選択した要素 P_i の P^{mix} を求める.

4. ルールの重み w を次式に従って更新する.

$$w(z,s,a)^{new} \leftarrow w(z,s,a)^{old} + w(z,s,a)^{old} \times P^{mix}(z,s,a) \quad (3.5)$$

重み更新後は利益共有法による方策学習を続ける.

第4章　クラスタリングによる要素選択

混合分布の構成要素となる同時分布数の増加に従い，混合分布に未知環境へ適応するための多様性を持たせることができると考えられるが，エージェントの設定や環境の規模によっては，混合分布の構成と方策改善に要する計算量が増大し，実環境を想定した場合には計算量の抑制は必要な問題となる．そこで，図 4.1 に示すように，統計的手法に基づいて混合分布の構成要素として適切な同時分布を選択できれば，計算量の抑制と，混合分布の方策改善への有効性を維持することが可能となる．本章では，同時分布の選択に導入するクラスタリングについて概要を述べ，適切なクラスタリング手法を検討する．

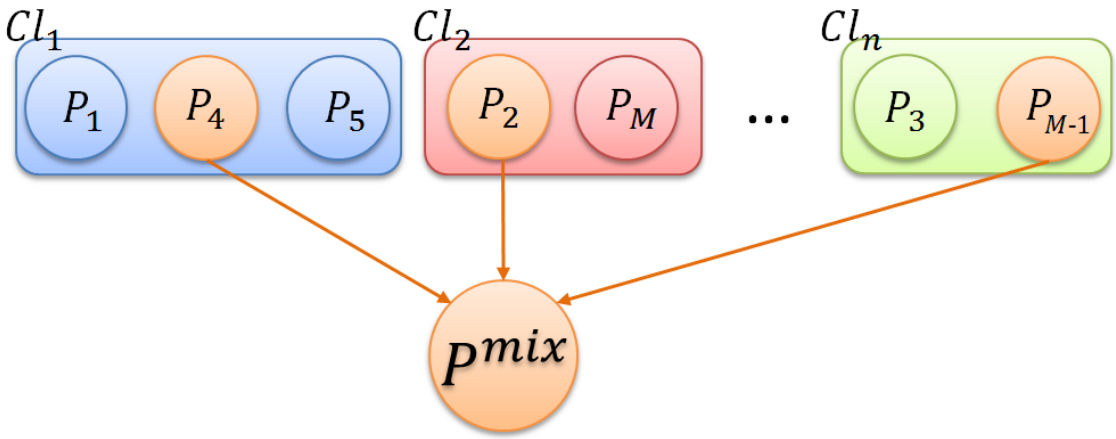


図 4.1: 構成要素の選択

4.1　クラスタリングとは

クラスタリングとは，分類対象となるデータ集合を，外的基準なしにクラスタと呼ばれる部分集合に切り分けて，似たものの同士で分類する教師なしデータ分類手法であり，機械学習やデータマイニング，パターン認識などの多くの分野で用いられる．さまざまな手法が提案されているが，大きく分けると階層的クラスタリングと非階層的クラスタリングの 2 つに分けられる [16, 17].

4.1.1　階層的クラスタリング

階層的クラスタリングは，1 個の対象データだけを含むそれぞれのクラスタ，つまり対象データと同数のクラスタがある初期状態から，クラスタ間の距離（非類似度）関数に基づいて，最小距離のクラスタを逐次的併合していく．最終的に全ての対象データが 1 つのクラスタに併合されるまで上記の操作を振り返すことで階層構造が獲得される．この階層構造は図 4.2 のような階層的なツリー（デンドログラム）によって表される．

階層的クラスタリングの代表的な手法としては，最短距離法，最長距離法、群平均法、ウォード法などがある [18].

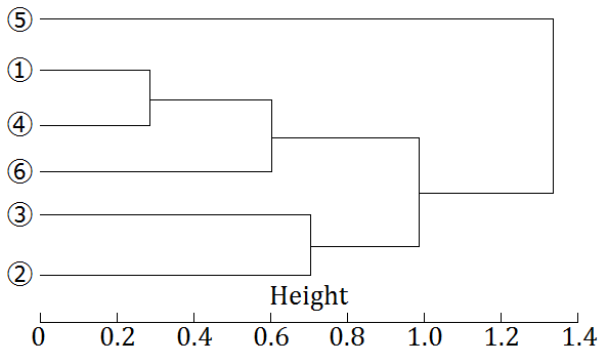


図 4.2: デンドログラムのサンプル

4.1.2　非階層的クラスタリング

非階層的クラスタリングは，分割の良さを示す評価関数をもとに，対象データの分類を振り返していき，評価関数を最適にする分割を探索する．

代表的手法である K-means 法は比較的簡単なアルゴリズムであることにより，さまざまな対応手法が提案され，現在広く用いられている．

4.2　採用する手法について

本研究では，クラスタリングの対象となるデータは確率分布であり，対象データ間の距離に Hellinger distance を用いるため，ウォード法や k-means 法などの数値ベクトルをもとにクラスタリングを行う手法は適用できない．また，階層的クラスタリングの最短距離法と最長距離法は，はずれ値などの影響を受けやすく，鎖効果と呼ばれる悪い判別が起こりやすい．よって，本研究では，最短距離法と最長距離法の中間的な性質を持ち鎖効果が起こらない群平均法を採用する．

群平均法では，クラスタ間の距離はそれぞれのクラスタに属する対象となるデータ間の全ての組み合わせの距離の平均値とする．

$$D(Cl_i, Cl_j) = \frac{1}{n_i n_j} \sum_{p_i \in Cl_i, p_j \in Cl_j} D(p_i, p_j) \tag{4.1}$$

n_i, n_j はそれぞれクラスタ Cl_i, Cl_j に含まれる対象データの数である．

本研究では，対象データは同時分布であり，距離関数 $D(p_i, p_j)$ は Hellinger Distance を用いる．

クラスタリング終了後，各クラスタに属する同時分布と未知環境で τ 回の学習の間に得られた同時分布間の D_H が最小のものを，混合分布の構成要素として選択する．適切なクラスタ数は問題設定などに応じて設定する．

4.3　分布の類似性の視覚化

分布間のそ相違度の尺度である D_H が距離の公理を待たすことから，クラスタリングの距離関数として導入したが，同時分布の関係を図示することで，視覚的に類似性を理解することも可能となる [3].

エージェントナビゲーション問題を適用例に，図 4.3 に示す環境 $E_1, \ldots, E_4$ で利益共有法による学習を実行する．ここでは実験設定と結果の詳細は省略する. 方策学習により各環境で得られた同時分布 $P_1, \ldots, P_4$ の間の Hellinger distance を表 4.1 に示す. この距離をもとに図示したものが図 4.4 であり，同時分布の類似性を直感的に理解することができる.

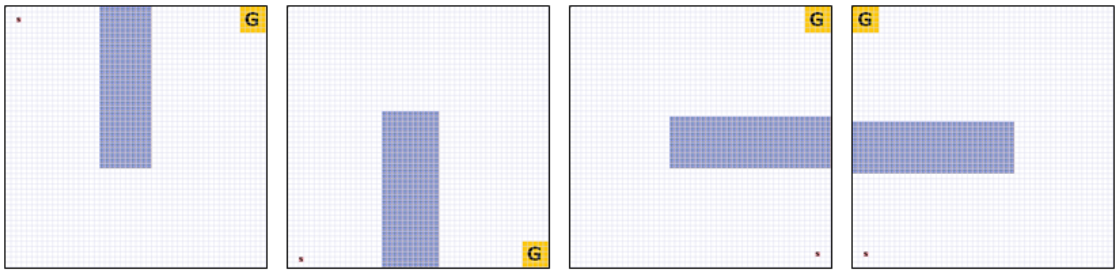


図 4.3: 環境 $E_1, \ldots, E_4$

表 4.1: Hellinger distance

	E_1	E_2	E_3	E_4
E_1	0			
E_2	1.141	0		
E_3	1.193	1.163	0	
E_4	1.310	1.121	1.331	0

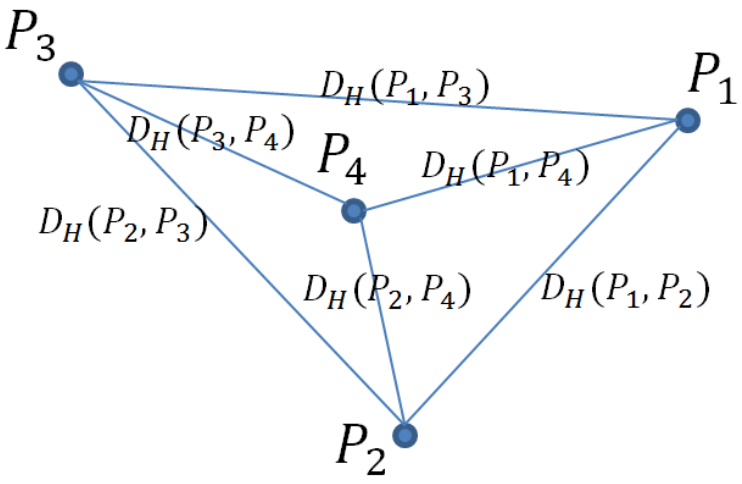


図 4.4: 視覚化のサンプル

第5章　本提案システムの流れ

図5.1に示すような強化学習エージェントにおける分布クラスタリングを用いた方策改善の流れは以下の手順となる．

1. エージェントが利益共有法により m 個の環境を学習する．
2. エージェントが過去に学習した m 個の環境に対応する同時分布に対して，クラスタ数 n 個でクラスタリングを行う．
3. クラスタリングにより選択された n 個の環境に対応する同時分布 $P_i(i = 1, \ldots, n)$ と，現在の環境（未知環境）で τ 回の方策学習の間に得られたサンプルの同時分布 Q の D_H を求める．

$$D_H(P_i, Q) = \left\{ \sum_x \left[P_i(x)^{\frac{1}{2}} - Q(x)^{\frac{1}{2}} \right]^2 \right\}^{\frac{1}{2}} \tag{5.1}$$

4. $D_H(P_i, Q)$ が最小の要素 P_i を各クラスタから選択する．
5. 最小 $D_H(P_i, Q)$ を利用し，混合パラメータ β_i を求める．

$$\beta_i = \frac{\sqrt{2} - D_H(P_i, Q)}{\sum_{j=1}^{n} (\sqrt{2} - D_{H_j})} \ (i = 1, \ldots, M) \tag{5.2}$$

6. 選択した要素 P_i の P^{mix} を求める．

$$P^{mix}(s, a) = \sum_{i=1}^{n} \beta_i P_i(s, a) \tag{5.3}$$

7. ルールの重み w を次式に従って更新する．

$$w(z, s, a)^{new} \leftarrow w(z, s, a)^{old} + w(z, s, a)^{old} \times P^{mix}(z, s, a) \tag{5.4}$$

そして，重み更新後は利益共有法による方策学習を続ける．

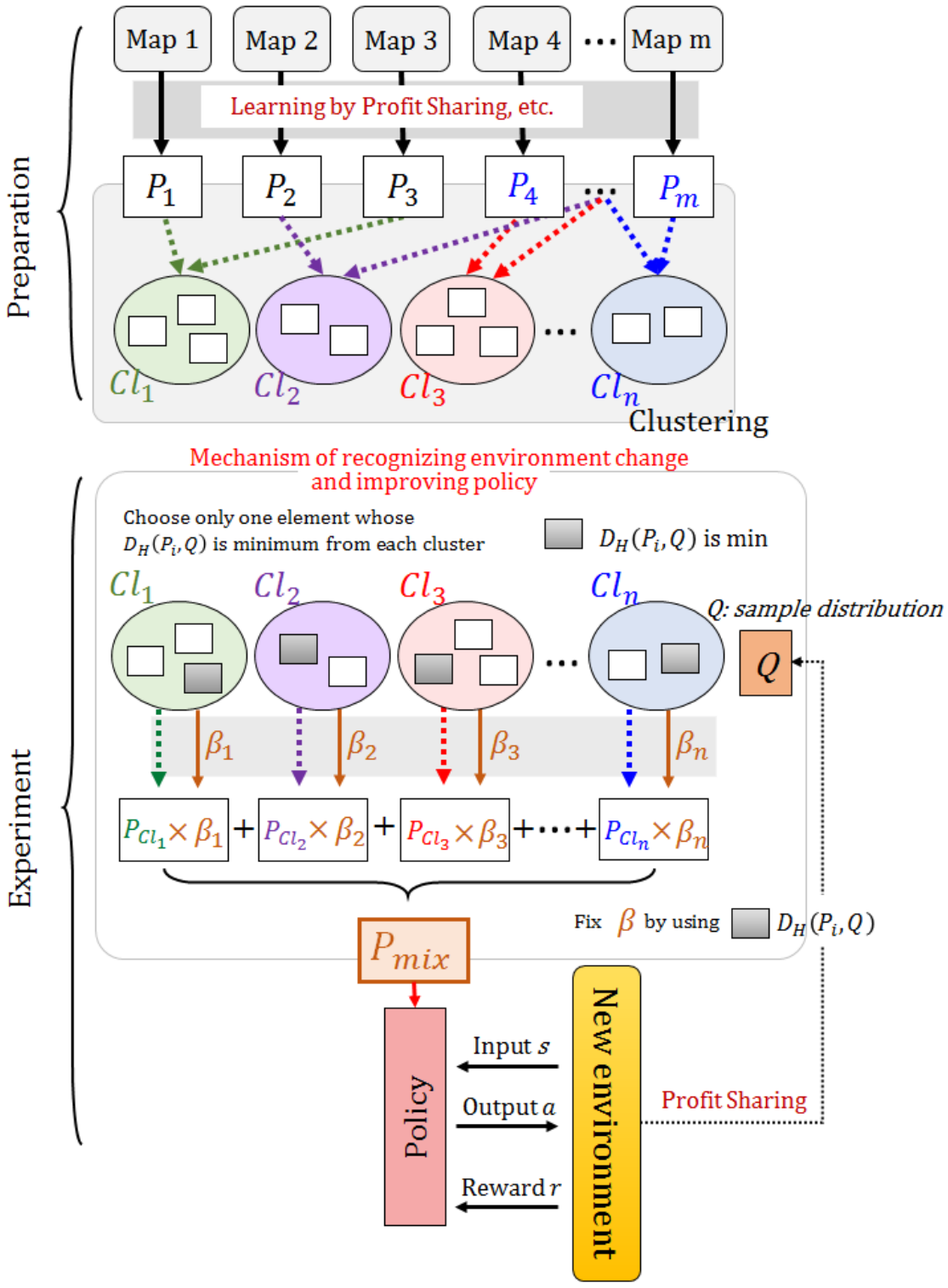

Preparation
Map 1
Map 2
Map 3
Map 4
Map m
Learning by Profit Sharing, etc.
P_1
P_2
P_3
P_4
P_m
Cl_1
Cl_2
Cl_3
Cl_n
Clustering
Experiment
Mechanism of recognizing environment change and improving policy
Choose only one element whose $D_H(P_i, Q)$ is minimum from each cluster
$D_H(P_i, Q)$ is min
Q: sample distribution
Q
β_1
β_2
β_3
β_n
$P_{Cl_1} \times \beta_1 + P_{Cl_2} \times \beta_2 + P_{Cl_3} \times \beta_3 + \cdots + P_{Cl_n} \times \beta_n$
Fix β by using $D_H(P_i, Q)$
P_{mix}
Policy
Input s
Output a
Reward r
New environment
Profit Sharing

図 5.1: システムの流れ

第6章　実験

本章では，混合分布とクラスタリングを用いた強化学習エージェントの方策改善の適用例として，エージェントナビゲーション問題に関する数字実験を行い，方策改善適用による未知環境への適用性能，混合分布の構成要素数の減少による混合分布の方策改善への有効性について評価する.

本実験では，方策改善適用による未知環境への適用性能，混合分布の構成要素数の減少による混合分布の方策改善への有効性について評価するため，まず50種類の環境で方策学習を行い，これによって得られた同時分布を混合分布の構成要素として，未知環境における方策学習において，分布クラスタリングを適用した混合分布による方策改善を行う.

6.1　エージェントナビゲーション問題

エージェントナビゲーション問題は，エージェントが置かれた環境の初期位置から目的地へと到達することを目的とする. 実験では図6.1のような環境で，障害物を避けながら出発点から目的地に到達すると報酬が与えられる.

1試行では，エージェントは行動選択回数は300 ×層数までとする. 行動選択回数以内に目的地にたどり着けば報酬rが与え，成功試行とする.そして，1試行終了後に行動選択回数は0にリセットされ，エージェントは初期位置から次の試行を再開する.

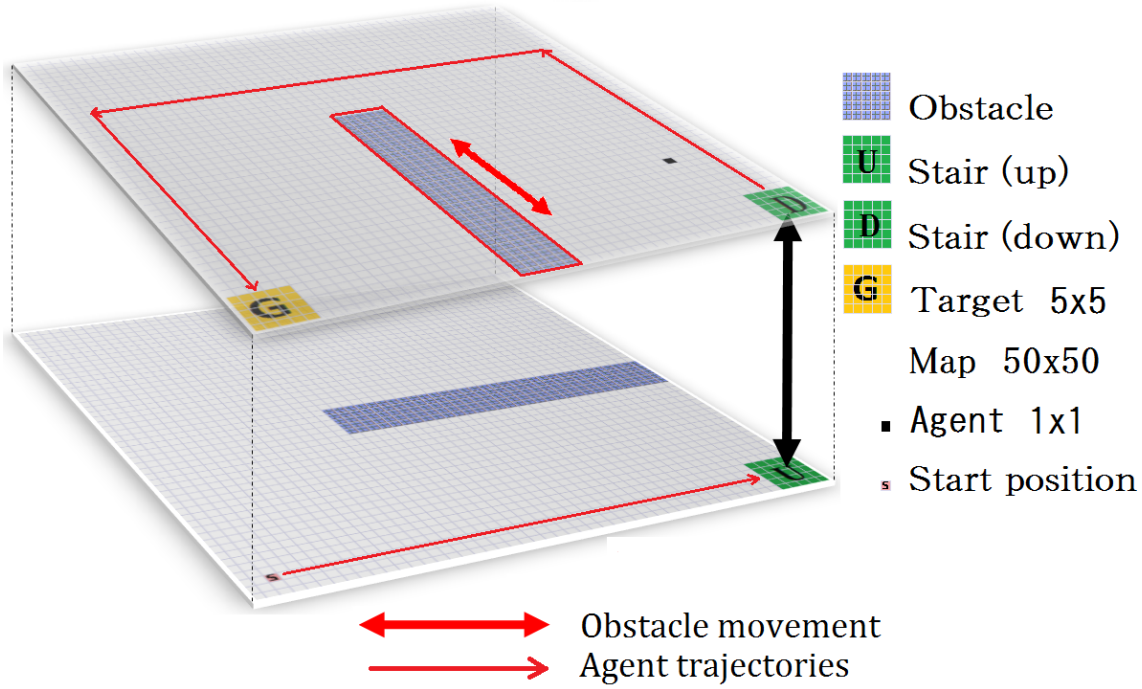


図 6.1: エージェントナビゲーション問題の環境

6.1.1 環境の設定

- エージェントの大きさ：1x1
- 目的地の大きさ：5x5
- 昇降：5x5
- 環境の大きさ：50x50

環境の違いは，エージェントの出発点，昇降，と目的地の位置，障害物の位置である．

6.1.2　エージェントの設定

エージェントは利益共有法によって方策を学習する．表 6.1 に示すようにエージェントの周りに障害物が存在する 8 方向 ($\uparrow, \downarrow, \leftarrow, \rightarrow, \nwarrow, \nearrow, \swarrow, \searrow$) の 256 通りに対応する整数値が入力として与えられ，表 6.2 に示す 8 方向 ($\uparrow, \downarrow, \leftarrow, \rightarrow, \nwarrow, \nearrow, \swarrow, \searrow$) の 8 通りの行動に対応する整数値を出力し，目的地に到達すると報酬 r が与えられる．つまり，エージェントは入力状態と出力行動の組み合わせとなる計 2048 個のルールを持つ．

表 6.1: 状態の種類の例.

障害物の位置とその値				
		 …	 …	
0	1	2	111	255

表 6.2: 移動の種類.

移動の方向			値		
$\nwarrow$	$\uparrow$	$\nearrow$	0	1	2
$\leftarrow$	Ⓐ	$\rightarrow$	3	Ⓐ	4
$\swarrow$	$\downarrow$	$\searrow$	5	6	7

6.1.3　混合モデルの構成要素

混合分布の構成用の 50 種類の既知環境を図 6.2 と図 6.3 に示す.

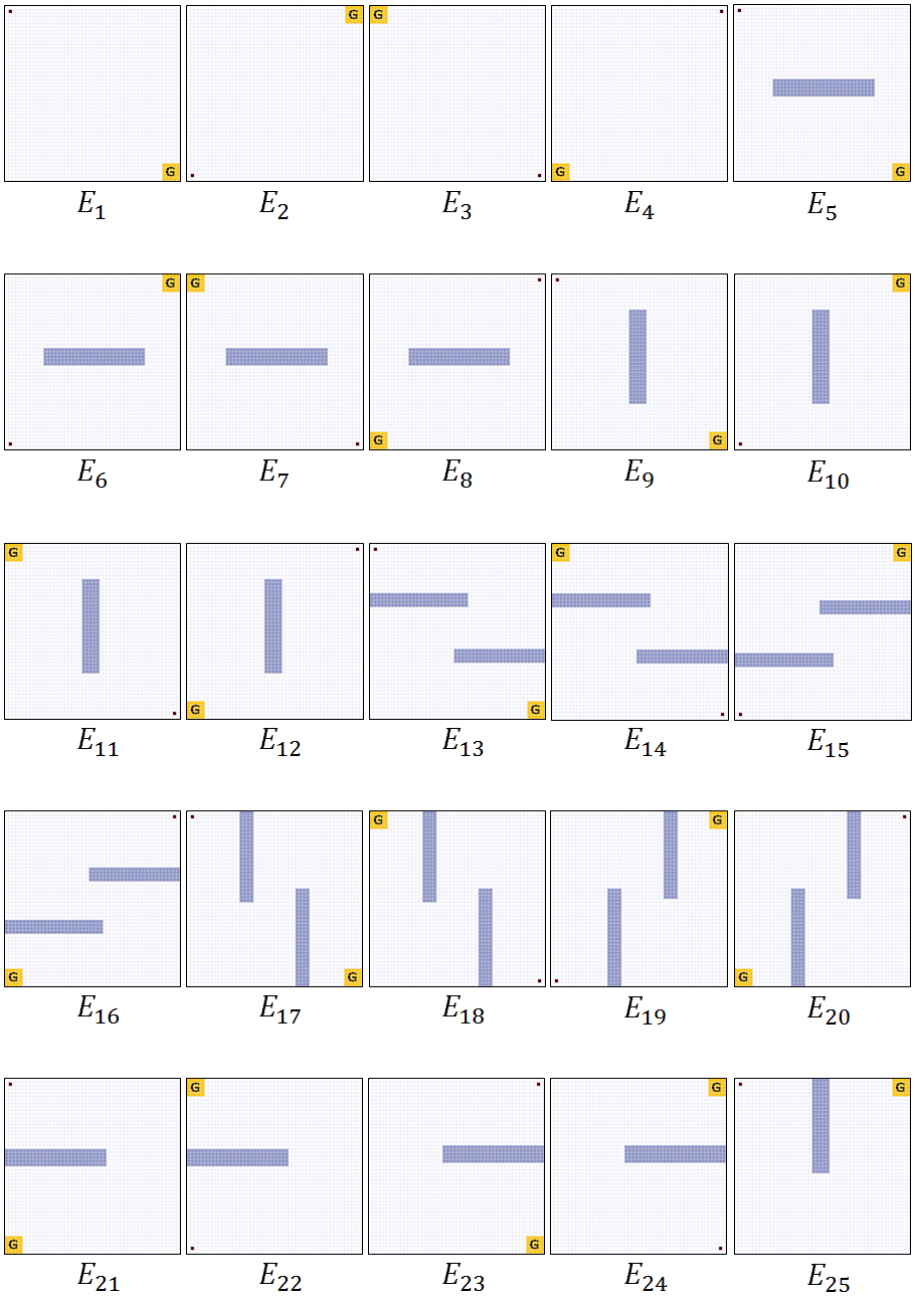


図 6.2: 混合分布構成ようの既知環境（E_1〜E_{25}）

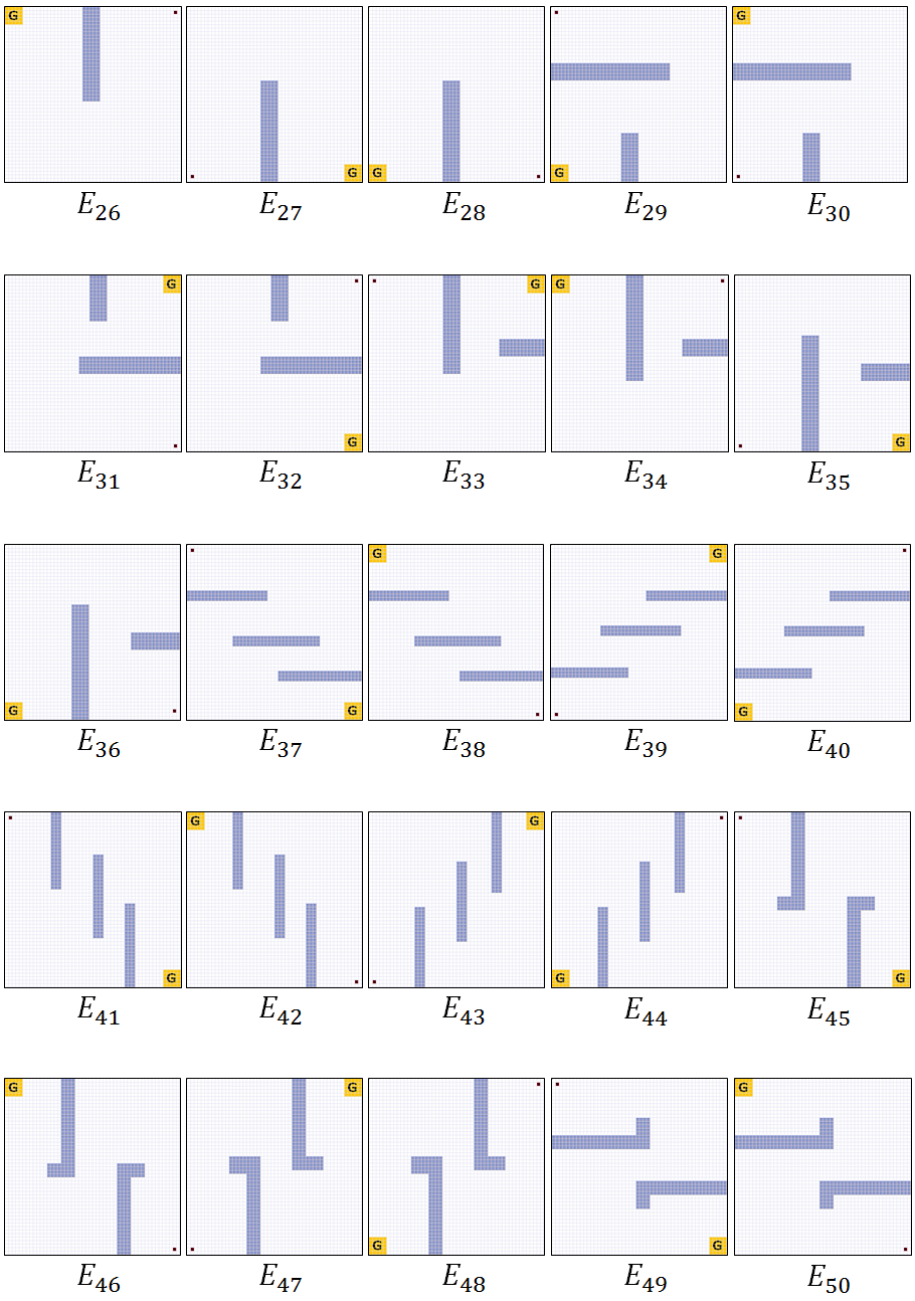


図 6.3: 混合分布構成ようの既知環境（E_{26}〜E_{50}）

6.1.4 クラスタリングの結果

$E_1, \ldots, E_{50}$ から得られた $P_1, \ldots, P_{50}$ 間の Hellinger distance を表 6.3 ～表 6.12 に示す．これらをもとに分布クラスタリングを行った結果のデンドログラムを図 6.4 に示す．これより 15，25 と 35 個のクラスタに属する環境が表 6.13 のように決定することができる．

表 6.3: 同時分布（P_1～P_{25}）と（P_1～P_{10}）間の Hellinger distance

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	0	1.3668	1.2713	1.3321	1.0241	1.3274	1.2374	1.3809	1.1152	1.1087
P_2	1.3668	0	1.1554	1.3799	1.1810	0.2252	1.3454	1.3910	1.1962	1.2588
P_3	1.2713	1.1554	0	1.2545	1.0780	1.1419	0.8838	1.2983	1.0163	0.9665
P_4	1.3321	1.3799	1.2545	0	1.1638	1.3600	1.2297	0.9191	1.0794	1.3173
P_5	1.0241	1.1810	1.0780	1.1638	0	1.1586	1.1572	1.2859	0.6364	0.9959
P_6	1.3274	0.2252	1.1419	1.3600	1.1586	0	1.3186	1.3773	1.1702	1.2438
P_7	1.2374	1.3454	0.8838	1.2297	1.1572	1.3186	0	1.2850	1.1099	1.1952
P_8	1.3809	1.3910	1.2983	0.9191	1.2859	1.3773	1.2850	0	1.2381	1.2208
P_9	1.1152	1.1962	1.0163	1.0794	0.6364	1.1702	1.1099	1.2381	0	1.0575
P_{10}	1.1087	1.2588	0.9665	1.3173	0.9959	1.2438	1.1952	1.2208	1.0575	0
P_{11}	1.2345	1.2705	0.7805	1.1966	1.0714	1.2495	0.6992	1.2575	1.0470	1.1305
P_{12}	1.2015	1.3097	1.1380	0.9063	1.0769	1.2950	1.0910	1.1861	0.9525	1.3050
P_{13}	1.1552	1.2137	1.1548	1.0577	0.7191	1.1927	1.2626	1.1649	0.7515	1.1083
P_{14}	1.2226	1.3295	0.7992	1.2269	1.1361	1.2833	0.7818	1.3664	1.0742	1.1816
P_{15}	1.3563	0.3300	1.1428	1.3319	1.1327	0.2878	1.2743	1.3334	1.1182	1.2420
P_{16}	1.3829	1.3815	1.3698	1.1401	1.2316	1.3785	1.3742	1.1069	1.3338	1.2857
P_{17}	1.0720	1.1788	1.0510	1.1195	0.5500	1.1476	1.1502	1.2995	0.6043	0.9951
P_{18}	1.2496	1.1393	0.7591	1.2065	1.0270	1.0858	0.8215	1.2739	0.9870	1.1756
P_{19}	1.2250	1.3684	1.0969	1.2393	1.1864	1.3434	1.1840	1.0807	1.1700	0.7542
P_{20}	1.2689	1.3507	1.1470	0.9653	1.1260	1.3227	1.0680	1.2022	1.0693	1.3356
P_{21}	0.7214	1.3659	1.2707	1.3056	0.9773	1.3434	1.2044	1.3590	1.0387	1.2009
P_{22}	1.2333	1.1179	0.7765	1.2213	1.0625	1.0698	0.7013	1.3107	1.0493	1.1576
P_{23}	1.3745	1.3785	1.3779	1.2514	1.1880	1.3756	1.3621	1.3701	1.2897	1.3358
P_{24}	1.3799	0.3243	1.1545	1.3882	1.2194	0.3357	1.3465	1.3479	1.2323	1.1536
P_{25}	1.1301	0.8464	1.1804	1.1363	1.0282	0.8225	1.2205	1.1800	0.9984	1.1976

表 6.4: 同時分布（P_{26}〜P_{50}）と（P_1〜P_{10}）間の Hellinger distance

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_{26}	1.3392	1.0491	0.9133	1.1653	1.1939	1.0406	1.2663	1.1935	1.1555	1.2122
P_{27}	1.3765	0.9844	1.1486	1.3858	1.0371	0.9866	1.3664	1.3857	1.0439	1.1771
P_{28}	1.3125	1.3693	1.3314	0.8143	1.1784	1.3670	1.3254	0.8232	1.1052	1.3194
P_{29}	1.0817	1.2779	1.1601	1.2383	0.7884	1.2595	1.1909	1.3687	0.8891	1.1789
P_{30}	1.2658	1.2518	0.6868	1.2776	1.1084	1.2620	0.8668	1.3795	1.0535	1.0402
P_{31}	1.3781	0.1922	1.1912	1.3888	1.1984	0.2188	1.3787	1.3932	1.2163	1.2840
P_{32}	1.3895	1.3781	1.3691	1.3777	1.2714	1.3717	1.3880	1.3581	1.3722	1.3484
P_{33}	0.8311	0.8886	1.1774	1.2811	1.0119	0.8477	1.2208	1.3307	1.0907	1.1294
P_{34}	1.3037	1.2279	0.7253	1.2049	1.1069	1.1945	1.0790	1.2240	1.0191	0.9848
P_{35}	1.3401	0.9892	1.1142	1.3235	0.9639	0.9829	1.3267	1.3821	0.9668	1.1579
P_{36}	1.3811	1.3626	1.3025	0.7099	1.2455	1.3617	1.3322	0.7266	1.1583	1.3021
P_{37}	0.5566	1.3339	1.2041	1.2672	0.9230	1.3011	1.1598	1.3224	1.0205	1.0752
P_{38}	1.2388	1.2648	0.9060	1.2146	1.1046	1.2298	0.6463	1.3529	1.0898	1.1723
P_{39}	1.3423	0.2604	1.1451	1.3785	1.1368	0.2489	1.3618	1.3841	1.1576	1.2120
P_{40}	1.3664	1.3783	1.3264	0.8217	1.2536	1.3781	1.3395	0.8291	1.1925	1.3121
P_{41}	1.0229	1.3023	1.1149	1.0080	0.8668	1.2783	1.1033	1.1680	0.7639	1.1133
P_{42}	1.2673	1.0652	0.8265	1.2412	1.0274	1.0685	0.8429	1.3191	1.0466	1.2320
P_{43}	1.1712	0.8130	1.2488	1.2685	1.1600	0.7850	1.3333	1.2899	1.1546	1.2179
P_{44}	1.3901	1.3937	1.3650	0.6436	1.2981	1.3870	1.3689	0.8136	1.2383	1.3617
P_{45}	1.0063	1.2992	1.1025	1.1759	0.7994	1.2868	1.0874	1.3269	0.7369	1.1115
P_{46}	1.2670	1.0475	0.9011	1.1887	1.0564	1.0382	0.9171	1.3661	1.0784	1.2729
P_{47}	1.1379	0.6187	1.0725	1.3340	1.0311	0.6251	1.2846	1.2981	1.0757	0.8807
P_{48}	1.3560	1.3456	1.2858	0.4982	1.1765	1.3356	1.2766	0.8236	1.1130	1.2994
P_{49}	0.7313	1.3099	1.1602	1.1329	0.8399	1.2859	1.1635	1.2552	0.8852	1.0118
P_{50}	1.2417	1.3040	0.8770	1.2397	1.1474	1.2640	0.7666	1.3427	1.1667	1.1788

表 6.5: 同時分布（P_1〜P_{25}）と（P_{11}〜P_{20}）間の Hellinger distance

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	1.2345	1.2015	1.1552	1.2226	1.3563	1.3829	1.0720	1.2496	1.2250	1.2689
P_2	1.2705	1.3097	1.2137	1.3295	0.3300	1.3815	1.1788	1.1393	1.3684	1.3507
P_3	0.7805	1.1380	1.1548	0.7992	1.1428	1.3698	1.0510	0.7591	1.0969	1.1470
P_4	1.1966	0.9063	1.0577	1.2269	1.3319	1.1401	1.1195	1.2065	1.2393	0.9653
P_5	1.0714	1.0769	0.7191	1.1361	1.1327	1.2316	0.5500	1.0270	1.1864	1.1260
P_6	1.2495	1.2950	1.1927	1.2833	0.2878	1.3785	1.1476	1.0858	1.3434	1.3227
P_7	0.6992	1.0910	1.2626	0.7818	1.2743	1.3742	1.1502	0.8215	1.1840	1.0680
P_8	1.2575	1.1861	1.1649	1.3664	1.3334	1.1069	1.2995	1.2739	1.0807	1.2022
P_9	1.0470	0.9525	0.7515	1.0742	1.1182	1.3338	0.6043	0.9870	1.1700	1.0693
P_{10}	1.1305	1.3050	1.1083	1.1816	1.2420	1.2857	0.9951	1.1756	0.7542	1.3356
P_{11}	0	1.0364	1.2060	0.7847	1.2322	1.3668	1.0487	0.6679	1.1300	1.0581
P_{12}	1.0364	0	1.0711	1.0807	1.2785	1.3638	0.9833	1.0791	1.2170	0.5612
P_{13}	1.2060	1.0711	0	1.2191	1.1621	1.1023	0.8054	1.1653	1.1662	1.1013
P_{14}	0.7847	1.0807	1.2191	0	1.2713	1.3721	1.1348	0.9450	1.2374	1.1030
P_{15}	1.2322	1.2785	1.1621	1.2713	0	1.3439	1.1234	1.0837	1.3090	1.3204
P_{16}	1.3668	1.3638	1.1023	1.3721	1.3439	0	1.2767	1.3656	1.2932	1.3718
P_{17}	1.0487	0.9833	0.8054	1.1348	1.1234	1.2767	0	1.0108	1.2170	1.0553
P_{18}	0.6679	1.0791	1.1653	0.9450	1.0837	1.3656	1.0108	0	1.2093	1.0411
P_{19}	1.1300	1.2170	1.1662	1.2374	1.3090	1.2932	1.2170	1.2093	0	1.1885
P_{20}	1.0581	0.5612	1.1013	1.1030	1.3204	1.3718	1.0553	1.0411	1.1885	0
P_{21}	1.2295	1.1615	1.1398	1.2204	1.3547	1.3710	1.0356	1.2096	1.2943	1.2439
P_{22}	0.7005	1.0963	1.1888	0.8018	1.0487	1.3746	1.0613	0.6431	1.2247	1.0877
P_{23}	1.3595	1.3243	1.2370	1.3667	1.3170	0.9922	1.2275	1.3439	1.3662	1.3790
P_{24}	1.2751	1.3418	1.2423	1.3714	0.4099	1.3637	1.2151	1.1454	1.2675	1.3750
P_{25}	1.2020	1.1653	0.9568	1.2437	0.7681	1.2796	1.0687	1.0804	1.1181	1.1661

表 6.6: 同時分布（P_{26}〜P_{50}）と（P_{11}〜P_{20}）間の Hellinger distance

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_{26}	1.1558	1.2255	1.1680	1.1421	1.0504	1.2845	1.1704	0.9639	1.2932	1.2286
P_{27}	1.2506	1.3195	1.0677	1.3580	1.0061	1.3936	1.0116	1.1626	1.3879	1.3874
P_{28}	1.3120	1.1402	1.0004	1.3087	1.3307	0.9805	1.2248	1.3249	1.2343	1.1925
P_{29}	1.1267	0.9935	0.9760	1.1647	1.2231	1.1478	0.7683	1.0950	1.3089	1.1458
P_{30}	0.8502	1.1552	1.1636	0.8828	1.2435	1.3697	1.0710	0.9930	1.1550	1.1468
P_{31}	1.3088	1.3279	1.2200	1.3717	0.3291	1.3798	1.1894	1.1564	1.3907	1.3617
P_{32}	1.3780	1.3715	1.2345	1.3701	1.3403	0.8201	1.2928	1.3743	1.3359	1.3742
P_{33}	1.2006	1.1761	1.1085	1.2120	0.8309	1.3443	1.0391	1.1252	1.1402	1.2156
P_{34}	1.0492	1.2019	1.1076	0.9203	1.1970	1.2690	1.0795	0.9625	1.1244	1.2212
P_{35}	1.1980	1.2456	1.0245	1.3175	1.0026	1.3928	0.9047	1.1126	1.3799	1.3171
P_{36}	1.3210	1.1535	1.0642	1.3139	1.3303	1.0849	1.2561	1.3366	1.2106	1.1777
P_{37}	1.1757	1.1728	1.0365	1.1474	1.3004	1.3131	0.9738	1.1675	1.1950	1.2295
P_{38}	0.6606	1.0650	1.2342	0.6505	1.1939	1.3796	1.1101	0.8555	1.1901	1.0788
P_{39}	1.2870	1.3278	1.1855	1.3350	0.3489	1.3696	1.1386	1.1372	1.3379	1.3730
P_{40}	1.3286	1.0781	1.1006	1.3329	1.3511	1.0996	1.2735	1.3422	1.2382	1.1551
P_{41}	1.0841	0.9066	0.7733	1.0968	1.2306	1.2549	0.9106	1.0708	1.0632	0.8745
P_{42}	0.7945	1.0794	1.1834	0.9582	1.0728	1.3750	1.0401	0.6748	1.2502	1.0552
P_{43}	1.3073	1.2479	1.1877	1.3224	0.7314	1.3795	1.2087	1.1999	1.0689	1.2251
P_{44}	1.3622	1.1026	1.1357	1.3761	1.3592	1.1231	1.3149	1.3658	1.2634	1.2074
P_{45}	1.0705	1.0645	0.9030	1.0882	1.2366	1.3205	0.8864	1.0851	1.1361	1.0344
P_{46}	0.8686	1.0571	1.1681	0.9236	1.0581	1.3813	1.0549	0.7476	1.3217	1.0189
P_{47}	1.2022	1.2737	1.1128	1.3033	0.6239	1.3358	1.0178	1.1242	1.0765	1.3023
P_{48}	1.2591	1.0331	1.0298	1.3038	1.2975	1.0096	1.1926	1.2361	1.2164	1.1069
P_{49}	1.1231	1.0275	0.8984	1.1474	1.2623	1.2810	0.8409	1.1314	1.1213	1.0701
P_{50}	0.6955	1.1262	1.2844	0.6754	1.2520	1.3790	1.1746	0.9047	1.1489	1.1164

表 6.7: 同時分布（P_1〜P_{25}）と（P_{21}〜P_{30}）間の Hellinger distance

	$P_2$1	$P_2$2	$P_2$3	$P_2$4	$P_2$5	$P_2$6	$P_2$7	$P_2$8	$P_2$9	P_{30}
P_1	0.7214	1.2333	1.3745	1.3799	1.1301	1.3392	1.3765	1.3125	1.0817	1.2658
P_2	1.3659	1.1179	1.3785	0.3243	0.8464	1.0491	0.9844	1.3693	1.2779	1.2518
P_3	1.2707	0.7765	1.3779	1.1545	1.1804	0.9133	1.1486	1.3314	1.1601	0.6868
P_4	1.3056	1.2213	1.2514	1.3882	1.1363	1.1653	1.3858	0.8143	1.2383	1.2776
P_5	0.9773	1.0625	1.1880	1.2194	1.0282	1.1939	1.0371	1.1784	0.7884	1.1084
P_6	1.3434	1.0698	1.3756	0.3357	0.8225	1.0406	0.9866	1.3670	1.2595	1.2620
P_7	1.2044	0.7013	1.3621	1.3465	1.2205	1.2663	1.3664	1.3254	1.1909	0.8668
P_8	1.3590	1.3107	1.3701	1.3479	1.1800	1.1935	1.3857	0.8232	1.3687	1.3795
P_9	1.0387	1.0493	1.2897	1.2323	0.9984	1.1555	1.0439	1.1052	0.8891	1.0535
P_{10}	1.2009	1.1576	1.3358	1.1536	1.1976	1.2122	1.1771	1.3194	1.1789	1.0402
P_{11}	1.2295	0.7005	1.3595	1.2751	1.2020	1.1558	1.2506	1.3120	1.1267	0.8502
P_{12}	1.1615	1.0963	1.3243	1.3418	1.1653	1.2255	1.3195	1.1402	0.9935	1.1552
P_{13}	1.1398	1.1888	1.2370	1.2423	0.9568	1.1680	1.0677	1.0004	0.9760	1.1636
P_{14}	1.2204	0.8018	1.3667	1.3714	1.2437	1.1421	1.3580	1.3087	1.1647	0.8828
P_{15}	1.3547	1.0487	1.3170	0.4099	0.7681	1.0504	1.0061	1.3307	1.2231	1.2435
P_{16}	1.3710	1.3746	0.9922	1.3637	1.2796	1.2845	1.3936	0.9805	1.1478	1.3697
P_{17}	1.0356	1.0613	1.2275	1.2151	1.0687	1.1704	1.0116	1.2248	0.7683	1.0710
P_{18}	1.2096	0.6431	1.3439	1.1454	1.0804	0.9639	1.1626	1.3249	1.0950	0.9930
P_{19}	1.2943	1.2247	1.3662	1.2675	1.1181	1.2932	1.3879	1.2343	1.3089	1.1550
P_{20}	1.2439	1.0877	1.3790	1.3750	1.1661	1.2286	1.3874	1.1925	1.1458	1.1468
P_{21}	0	1.2234	1.3811	1.3854	1.1644	1.3135	1.3743	1.2475	1.0462	1.2247
P_{22}	1.2234	0	1.3645	1.1122	1.0856	1.0301	1.1646	1.3513	1.1088	0.9004
P_{23}	1.3811	1.3645	0	1.3818	1.3163	1.3270	1.3928	1.3233	1.2372	1.3641
P_{24}	1.3854	1.1122	1.3818	0	0.8831	1.0611	0.9913	1.3889	1.2982	1.2457
P_{25}	1.1644	1.0856	1.3163	0.8831	0	1.0575	1.2561	1.1179	1.2139	1.2304

表 6.8: 同時分布（P_{26}〜P_{50}）と（P_{21}〜P_{30}）間の Hellinger distance

	P_21	P_22	P_23	P_24	P_25	P_26	P_27	P_28	P_29	P_{30}
P_{26}	1.3135	1.0301	1.3270	1.0611	1.0575	0	1.1313	1.1897	1.2497	1.0055
P_{27}	1.3743	1.1646	1.3928	0.9913	1.2561	1.1313	0	1.3843	1.1831	1.1906
P_{28}	1.2475	1.3513	1.3233	1.3889	1.1179	1.1897	1.3843	0	1.2647	1.3116
P_{29}	1.0462	1.1088	1.2372	1.2982	1.2139	1.2497	1.1831	1.2647	0	1.1880
P_{30}	1.2247	0.9004	1.3641	1.2457	1.2304	1.0055	1.1906	1.3116	1.1880	0
P_{31}	1.3720	1.1365	1.3757	0.2654	0.8556	1.0511	0.9761	1.3842	1.2888	1.2880
P_{32}	1.3950	1.3846	0.7234	1.3766	1.3492	1.3502	1.3938	1.3892	1.2502	1.3816
P_{33}	0.9497	1.0927	1.3166	0.9144	0.6054	1.1601	1.3232	1.2840	1.1184	1.2228
P_{34}	1.2657	1.0171	1.2884	1.2257	1.1828	0.8066	1.1402	1.1913	1.2039	0.8925
P_{35}	1.3380	1.1168	1.3917	0.9997	1.2338	1.1305	0.1844	1.3837	1.1181	1.1604
P_{36}	1.3491	1.3506	1.3901	1.3823	1.1308	1.1728	1.3700	0.4829	1.3610	1.3306
P_{37}	0.6309	1.1772	1.3538	1.3568	1.0503	1.2782	1.3194	1.2394	1.0180	1.2250
P_{38}	1.2254	0.6138	1.3657	1.2750	1.1865	1.1934	1.3007	1.3431	1.1504	0.8862
P_{39}	1.3317	1.1163	1.3744	0.2693	0.8489	1.0315	0.9823	1.3689	1.2436	1.2580
P_{40}	1.3328	1.3538	1.3305	1.3877	1.1523	1.1478	1.3916	0.6549	1.3606	1.3260
P_{41}	1.0902	1.1126	1.2758	1.3433	0.8711	1.2430	1.2833	1.0449	1.0646	1.1379
P_{42}	1.2360	0.5248	1.3744	1.0951	1.0776	0.9830	1.1327	1.3345	1.0982	0.9261
P_{43}	1.2466	1.2089	1.3165	0.8289	0.5290	1.1463	1.3439	1.2445	1.2961	1.3134
P_{44}	1.3649	1.3744	1.2902	1.3959	1.1735	1.1524	1.3890	0.6916	1.3794	1.3786
P_{45}	1.0534	1.1250	1.1997	1.3362	0.9762	1.3036	1.2765	1.1969	1.1128	1.0894
P_{46}	1.2594	0.5595	1.3897	1.1026	1.0828	1.0216	1.1317	1.3339	1.0851	1.0049
P_{47}	1.2011	1.1192	1.3408	0.5551	0.7242	1.0613	1.0926	1.3068	1.1910	1.1326
P_{48}	1.3266	1.2699	1.1337	1.3546	1.1108	1.1144	1.3257	0.7140	1.2497	1.2885
P_{49}	0.8628	1.1478	1.3306	1.3414	0.9718	1.2500	1.2901	1.1932	0.8796	1.1795
P_{50}	1.2164	0.6507	1.3688	1.3205	1.2329	1.1839	1.3778	1.3366	1.1706	0.9774

表 6.9: 同時分布（P_1〜P_{25}）と（P_{31}〜P_{40}）間の Hellinger distance

	P_31	P_32	P_33	P_34	P_35	P_36	P_37	P_38	P_39	P_{40}
P_1	1.3781	1.3895	0.8311	1.3037	1.3401	1.3811	0.5566	1.2388	1.3423	1.3664
P_2	0.1922	1.3781	0.8886	1.2279	0.9892	1.3626	1.3339	1.2648	0.2604	1.3783
P_3	1.1912	1.3691	1.1774	0.7253	1.1142	1.3025	1.2041	0.9060	1.1451	1.3264
P_4	1.3888	1.3777	1.2811	1.2049	1.3235	0.7099	1.2672	1.2146	1.3785	0.8217
P_5	1.1984	1.2714	1.0119	1.1069	0.9639	1.2455	0.9230	1.1046	1.1368	1.2536
P_6	0.2188	1.3717	0.8477	1.1945	0.9829	1.3617	1.3011	1.2298	0.2489	1.3781
P_7	1.3787	1.3880	1.2208	1.0790	1.3267	1.3322	1.1598	0.6463	1.3618	1.3395
P_8	1.3932	1.3581	1.3307	1.2240	1.3821	0.7266	1.3224	1.3529	1.3841	0.8291
P_9	1.2163	1.3722	1.0907	1.0191	0.9668	1.1583	1.0205	1.0898	1.1576	1.1925
P_{10}	1.2840	1.3484	1.1294	0.9848	1.1579	1.3021	1.0752	1.1723	1.2120	1.3121
P_{11}	1.3088	1.3780	1.2006	1.0492	1.1980	1.3210	1.1757	0.6606	1.2870	1.3286
P_{12}	1.3279	1.3715	1.1761	1.2019	1.2456	1.1535	1.1728	1.0650	1.3278	1.0781
P_{13}	1.2200	1.2345	1.1085	1.1076	1.0245	1.0642	1.0365	1.2342	1.1855	1.1006
P_{14}	1.3717	1.3701	1.2120	0.9203	1.3175	1.3139	1.1474	0.6505	1.3350	1.3329
P_{15}	0.3291	1.3403	0.8309	1.1970	1.0026	1.3303	1.3004	1.1939	0.3489	1.3511
P_{16}	1.3798	0.8201	1.3443	1.2690	1.3928	1.0849	1.3131	1.3796	1.3696	1.0996
P_{17}	1.1894	1.2928	1.0391	1.0795	0.9047	1.2561	0.9738	1.1101	1.1386	1.2735
P_{18}	1.1564	1.3743	1.1252	0.9625	1.1126	1.3366	1.1675	0.8555	1.1372	1.3422
P_{19}	1.3907	1.3359	1.1402	1.1244	1.3799	1.2106	1.1950	1.1901	1.3379	1.2382
P_{20}	1.3617	1.3742	1.2156	1.2212	1.3171	1.1777	1.2295	1.0788	1.3730	1.1551
P_{21}	1.3720	1.3950	0.9497	1.2657	1.3380	1.3491	0.6309	1.2254	1.3317	1.3328
P_{22}	1.1365	1.3846	1.0927	1.0171	1.1168	1.3506	1.1772	0.6138	1.1163	1.3538
P_{23}	1.3757	0.7234	1.3166	1.2884	1.3917	1.3901	1.3538	1.3657	1.3744	1.3305
P_{24}	0.2654	1.3766	0.9144	1.2257	0.9997	1.3823	1.3568	1.2750	0.2693	1.3877
P_{25}	0.8556	1.3492	0.6054	1.1828	1.2338	1.1308	1.0503	1.1865	0.8489	1.1523

表 6.10: 同時分布（P_{26}〜P_{50}）と（P_{31}〜P_{40}）間の Hellinger distance

	P_31	P_32	P_33	P_34	P_35	P_36	P_37	P_38	P_39	P_{40}
P_{26}	1.0511	1.3502	1.1601	0.8066	1.1305	1.1728	1.2782	1.1934	1.0315	1.1478
P_{27}	0.9761	1.3938	1.3232	1.1402	0.1844	1.3700	1.3194	1.3007	0.9823	1.3916
P_{28}	1.3842	1.3892	1.2840	1.1913	1.3837	0.4829	1.2394	1.3431	1.3689	0.6549
P_{29}	1.2888	1.2502	1.1184	1.2039	1.1181	1.3610	1.0180	1.1504	1.2436	1.3606
P_{30}	1.2880	1.3816	1.2228	0.8925	1.1604	1.3306	1.2250	0.8862	1.2580	1.3260
P_{31}	0	1.3728	0.8972	1.2527	0.9826	1.3833	1.3468	1.3028	0.1827	1.3917
P_{32}	1.3728	0	1.3461	1.2987	1.3934	1.3924	1.3596	1.3788	1.3651	1.3515
P_{33}	0.8972	1.3461	0	1.2583	1.2867	1.3364	0.8479	1.1833	0.8726	1.3206
P_{34}	1.2527	1.2987	1.2583	0	1.1336	1.1887	1.2189	1.0645	1.2031	1.2050
P_{35}	0.9826	1.3934	1.2867	1.1336	0	1.3668	1.2793	1.2551	0.9810	1.3864
P_{36}	1.3833	1.3924	1.3364	1.1887	1.3668	0	1.3082	1.3419	1.3760	0.5549
P_{37}	1.3468	1.3596	0.8479	1.2189	1.2793	1.3082	0	1.1871	1.3077	1.3100
P_{38}	1.3028	1.3788	1.1833	1.0645	1.2551	1.3419	1.1871	0	1.2744	1.3438
P_{39}	0.1827	1.3651	0.8726	1.2031	0.9810	1.3760	1.3077	1.2744	0	1.3859
P_{40}	1.3917	1.3515	1.3206	1.2050	1.3864	0.5549	1.3100	1.3438	1.3859	0
P_{41}	1.3276	1.3398	0.9936	1.1636	1.2310	1.0853	0.9231	1.1100	1.3207	1.1112
P_{42}	1.0941	1.3851	1.1200	1.0317	1.0909	1.3487	1.2057	0.8084	1.0847	1.3517
P_{43}	0.8147	1.3764	0.5677	1.2736	1.3326	1.2601	1.1473	1.2958	0.8104	1.2670
P_{44}	1.3962	1.3907	1.3539	1.2323	1.3887	0.5499	1.3332	1.3793	1.3943	0.4410
P_{45}	1.3183	1.3399	0.9986	1.2295	1.2197	1.2942	0.9295	1.1211	1.3019	1.2831
P_{46}	1.0902	1.3959	1.0987	1.0816	1.0698	1.3528	1.2118	0.7895	1.0844	1.3510
P_{47}	0.6275	1.3776	0.6492	1.1592	1.0786	1.3300	1.1070	1.2490	0.5882	1.3288
P_{48}	1.3585	1.2710	1.2964	1.1295	1.3098	0.6396	1.2849	1.2731	1.3540	0.7541
P_{49}	1.3350	1.3304	0.8672	1.2000	1.2320	1.2059	0.5523	1.1470	1.3062	1.2244
P_{50}	1.3489	1.3697	1.1892	1.0614	1.3387	1.3402	1.1590	0.6309	1.3143	1.3554

表 6.11: 同時分布（P_1〜P_{25}）と（P_{41}〜P_{50}）間の Hellinger distance

	P_41	P_42	P_43	P_44	P_45	P_46	P_47	P_48	P_49	P_{50}
P_1	1.0229	1.2673	1.1712	1.3901	1.0063	1.2670	1.1379	1.3560	0.7313	1.2417
P_2	1.3023	1.0652	0.8130	1.3937	1.2992	1.0475	0.6187	1.3456	1.3099	1.3040
P_3	1.1149	0.8265	1.2488	1.3650	1.1025	0.9011	1.0725	1.2858	1.1602	0.8770
P_4	1.0080	1.2412	1.2685	0.6436	1.1759	1.1887	1.3340	0.4982	1.1329	1.2397
P_5	0.8668	1.0274	1.1600	1.2981	0.7994	1.0564	1.0311	1.1765	0.8399	1.1474
P_6	1.2783	1.0685	0.7850	1.3870	1.2868	1.0382	0.6251	1.3356	1.2859	1.2640
P_7	1.1033	0.8429	1.3333	1.3689	1.0874	0.9171	1.2846	1.2766	1.1635	0.7666
P_8	1.1680	1.3191	1.2899	0.8136	1.3269	1.3661	1.2981	0.8236	1.2552	1.3427
P_9	0.7639	1.0466	1.1546	1.2383	0.7369	1.0784	1.0757	1.1130	0.8852	1.1667
P_{10}	1.1133	1.2320	1.2179	1.3617	1.1115	1.2729	0.8807	1.2994	1.0118	1.1788
P_{11}	1.0841	0.7945	1.3073	1.3622	1.0705	0.8686	1.2022	1.2591	1.1231	0.6955
P_{12}	0.9066	1.0794	1.2479	1.1026	1.0645	1.0571	1.2737	1.0331	1.0275	1.1262
P_{13}	0.7733	1.1834	1.1877	1.1357	0.9030	1.1681	1.1128	1.0298	0.8984	1.2844
P_{14}	1.0968	0.9582	1.3224	1.3761	1.0882	0.9236	1.3033	1.3038	1.1474	0.6754
P_{15}	1.2306	1.0728	0.7314	1.3592	1.2366	1.0581	0.6239	1.2975	1.2623	1.2520
P_{16}	1.2549	1.3750	1.3795	1.1231	1.3205	1.3813	1.3358	1.0096	1.2810	1.3790
P_{17}	0.9106	1.0401	1.2087	1.3149	0.8864	1.0549	1.0178	1.1926	0.8409	1.1746
P_{18}	1.0708	0.6748	1.1999	1.3658	1.0851	0.7476	1.1242	1.2361	1.1314	0.9047
P_{19}	1.0632	1.2502	1.0689	1.2634	1.1361	1.3217	1.0765	1.2164	1.1213	1.1489
P_{20}	0.8745	1.0552	1.2251	1.2074	1.0344	1.0189	1.3023	1.1069	1.0701	1.1164
P_{21}	1.0902	1.2360	1.2466	1.3649	1.0534	1.2594	1.2011	1.3266	0.8628	1.2164
P_{22}	1.1126	0.5248	1.2089	1.3744	1.1250	0.5595	1.1192	1.2699	1.1478	0.6507
P_{23}	1.2758	1.3744	1.3165	1.2902	1.1997	1.3897	1.3408	1.1337	1.3306	1.3688
P_{24}	1.3433	1.0951	0.8289	1.3959	1.3362	1.1026	0.5551	1.3546	1.3414	1.3205
P_{25}	0.8711	1.0776	0.5290	1.1735	0.9762	1.0828	0.7242	1.1108	0.9718	1.2329

表 6.12: 同時分布（P_{26}〜P_{50}）と（P_{41}〜P_{50}）間の Hellinger distance

	P_41	P_42	P_43	P_44	P_45	P_46	P_47	P_48	P_49	P_{50}
P_{26}	1.2430	0.9830	1.1463	1.1524	1.3036	1.0216	1.0613	1.1144	1.2500	1.1839
P_{27}	1.2833	1.1327	1.3439	1.3890	1.2765	1.1317	1.0926	1.3257	1.2901	1.3778
P_{28}	1.0449	1.3345	1.2445	0.6916	1.1969	1.3339	1.3068	0.7140	1.1932	1.3366
P_{29}	1.0646	1.0982	1.2961	1.3794	1.1128	1.0851	1.1910	1.2497	0.8796	1.1706
P_{30}	1.1379	0.9261	1.3134	1.3786	1.0894	1.0049	1.1326	1.2885	1.1795	0.9774
P_{31}	1.3276	1.0941	0.8147	1.3962	1.3183	1.0902	0.6275	1.3585	1.3350	1.3489
P_{32}	1.3398	1.3851	1.3764	1.3907	1.3399	1.3959	1.3776	1.2710	1.3304	1.3697
P_{33}	0.9936	1.1200	0.5677	1.3539	0.9986	1.0987	0.6492	1.2964	0.8672	1.1892
P_{34}	1.1636	1.0317	1.2736	1.2323	1.2295	1.0816	1.1592	1.1295	1.2000	1.0614
P_{35}	1.2310	1.0909	1.3326	1.3887	1.2197	1.0698	1.0786	1.3098	1.2320	1.3387
P_{36}	1.0853	1.3487	1.2601	0.5499	1.2942	1.3528	1.3300	0.6396	1.2059	1.3402
P_{37}	0.9231	1.2057	1.1473	1.3332	0.9295	1.2118	1.1070	1.2849	0.5523	1.1590
P_{38}	1.1100	0.8084	1.2958	1.3793	1.1211	0.7895	1.2490	1.2731	1.1470	0.6309
P_{39}	1.3207	1.0847	0.8104	1.3943	1.3019	1.0844	0.5882	1.3540	1.3062	1.3143
P_{40}	1.1112	1.3517	1.2670	0.4410	1.2831	1.3510	1.3288	0.7541	1.2244	1.3554
P_{41}	0	1.1291	1.0317	1.1572	0.5368	1.1182	1.1090	1.0378	0.6468	1.1589
P_{42}	1.1291	0	1.1945	1.3715	1.1161	0.4385	1.1119	1.2653	1.1717	0.8655
P_{43}	1.0317	1.1945	0	1.2858	1.0587	1.1948	0.6439	1.2511	1.0843	1.2953
P_{44}	1.1572	1.3715	1.2858	0	1.3396	1.3829	1.3657	0.5552	1.2518	1.3842
P_{45}	0.5368	1.1161	1.0587	1.3396	0	1.1422	1.0868	1.2344	0.7966	1.1725
P_{46}	1.1182	0.4385	1.1948	1.3829	1.1422	0	1.1258	1.2674	1.1572	0.8532
P_{47}	1.1090	1.1119	0.6439	1.3657	1.0868	1.1258	0	1.3088	1.0556	1.2879
P_{48}	1.0378	1.2653	1.2511	0.5552	1.2344	1.2674	1.3088	0	1.1637	1.3066
P_{49}	0.6468	1.1717	1.0843	1.2518	0.7966	1.1572	1.0556	1.1637	0	1.1668
P_{50}	1.1589	0.8655	1.2953	1.3842	1.1725	0.8532	1.2879	1.3066	1.1668	0

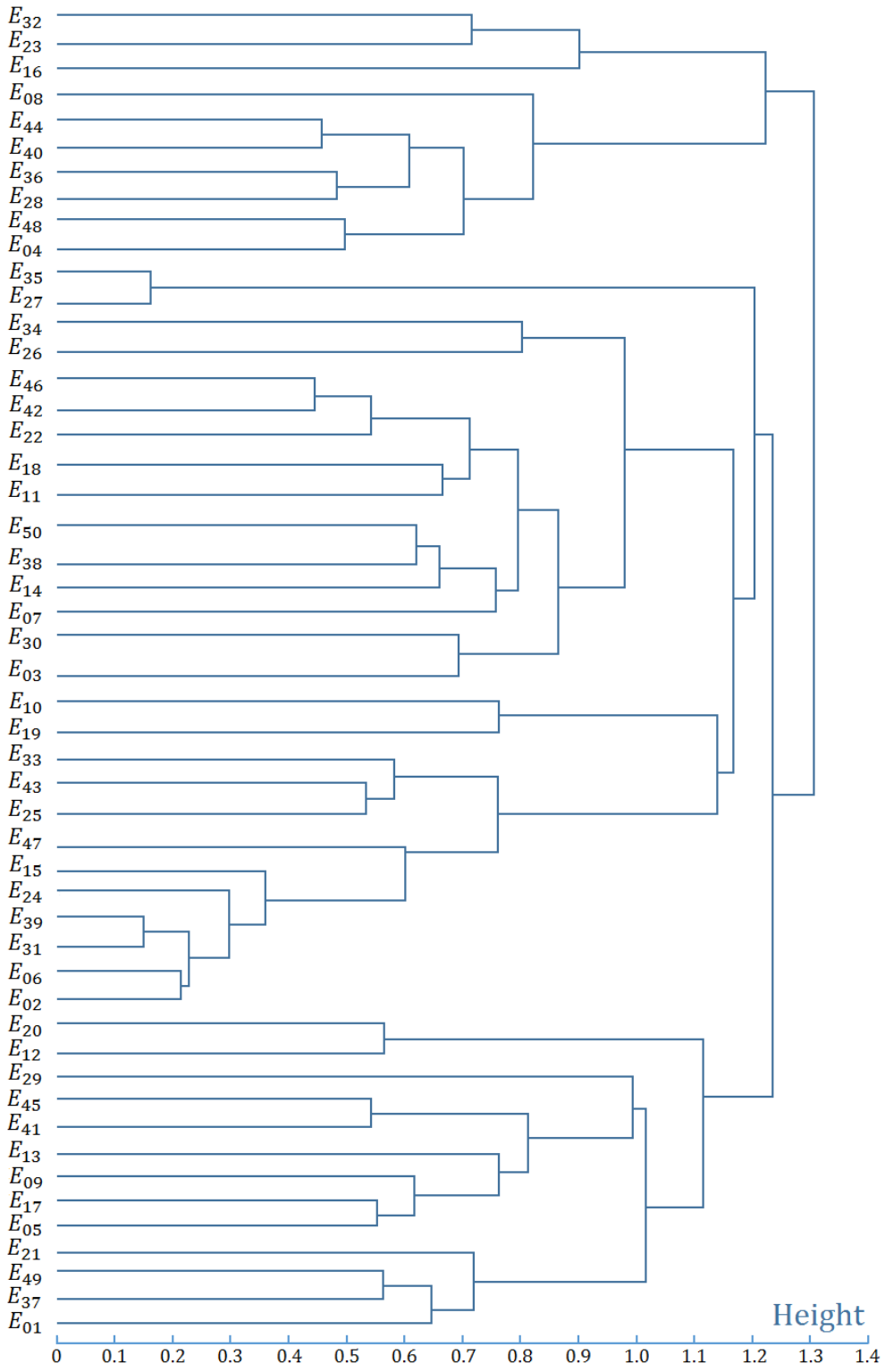


図 6.4: 分布クラスタリングのデンドログラム

表 6.13: クラスタ数 n および各クラスタに含まれる環境

n	クラスタ Cl_i に含まれる環境
15	$Cl_1 = \{E_1, E_{37}, E_{49}, E_{21}\}$, $Cl_2 = \{E_5, E_{17}, E_9, E_{13}\}$, $Cl_3 = \{E_{41}, E_{45}\}$, $Cl_4 = \{E_{29}\}$, $Cl_5 = \{E_{12}, E_{20}\}$, $Cl_6 = \{E_2, E_6, E_{31}, E_{39}, E_{24}, E_{15}, E_{47}, E_{25}, E_{43}, E_{33}\}$, $Cl_7 = \{E_{19}, E_{10}\}$, $Cl_8 = \{E_3, E_{30}\}$, $Cl_9 = \{E_7, E_{14}, E_{38}, E_{50}, E_{11}, E_{18}, E_{22}, E_{42}, E_{46}\}$, $Cl_{10} = \{E_{26}, E_{34}\}$, $Cl_{11} = \{E_{27}, E_{35}\}$, $Cl_{12} = \{E_4, E_{48}, E_{28}, E_{36}, E_{40}, E_{44}\}$, $Cl_{13} = \{E_8\}$, $Cl_{14} = \{E_{16}\}$, $Cl_{15} = \{E_{23}, E_{32}\}$
25	$Cl_1 = \{E_1, E_{37}, E_{49}\}$, $Cl_2 = \{E_{21}\}$, $Cl_3 = \{E_5, E_{17}, E_9\}$, $Cl_4 = \{E_{13}\}$, $Cl_5 = \{E_{41}, E_{45}\}$, $Cl_6 = \{E_{29}\}$, $Cl_7 = \{E_{12}, E_{20}\}$, $Cl_8 = \{E_2, E_6, E_{31}, E_{39}, E_{24}, E_{15}, E_{47}\}$, $Cl_9 = \{E_{25}, E_{43}, E_{33}\}$, $Cl_{10} = \{E_{19}\}$, $Cl_{11} = \{E_{10}\}$, $Cl_{12} = \{E_3, E_{30}\}$, $Cl_{13} = \{E_7\}$, $Cl_{14} = \{E_{14}, E_{38}, E_{50}\}$, $Cl_{15} = \{E_{11}, E_{18}\}$, $Cl_{16} = \{E_{22}, E_{42}, E_{46}\}$, $Cl_{17} = \{E_{26}\}$, $Cl_{18} = \{E_{34}\}$, $Cl_{19} = \{E_{27}, E_{35}\}$, $Cl_{20} = \{E_4, E_{48}\}$, $Cl_{21} = \{E_{28}, E_{36}, E_{40}, E_{44}\}$, $Cl_{22} = \{E_8\}$, $Cl_{23} = \{E_{16}\}$, $Cl_{24} = \{E_{23}\}$, $Cl_{25} = \{E_{32}\}$
35	$Cl_1 = \{E_1\}$, $Cl_2 = \{E_{37}, E_{49}\}$, $Cl_3 = \{E_{21}\}$, $Cl_4 = \{E_5, E_{17}\}$, $Cl_5 = \{E_9\}$, $Cl_6 = \{E_{13}\}$, $Cl_7 = \{E_{41}, E_{45}\}$, $Cl_8 = \{E_{29}\}$, $Cl_9 = \{E_{12}, \}$, $Cl_{10} = \{E_{20}\}$, $Cl_{11} = \{E_2, E_6, E_{31}, E_{39}, E_{24}, E_{15}\}$, $Cl_{12} = \{E_{47}\}$, $Cl_{13} = \{E_{25}, E_{43}\}$, $Cl_{14} = \{E_{33}\}$, $Cl_{15} = \{E_{19}\}$, $Cl_{16} = \{E_{10}\}$, $Cl_{17} = \{E_3\}$, $Cl_{18} = \{E_{30}\}$, $Cl_{19} = \{E_7\}$, $Cl_{20} = \{E_{14}\}$, $Cl_{21} = \{E_{38}\}$, $Cl_{22} = \{E_{50}\}$, $Cl_{23} = \{E_{11}\}$, $Cl_{24} = \{E_{18}\}$, $Cl_{25} = \{E_{22}, E_{42}, E_{46}\}$, $Cl_{26} = \{E_{26}\}$, $Cl_{27} = \{E_{34}\}$, $Cl_{28} = \{E_{27}, E_{35}\}$, $Cl_{29} = \{E_4, E_{48}\}$, $Cl_{30} = \{E_{28}, E_{36}\}$, $Cl_{31} = \{E_{40}, E_{44}\}$, $Cl_{32} = \{E_8\}$, $Cl_{33} = \{E_{16}\}$, $Cl_{34} = \{E_{23}\}$, $Cl_{35} = \{E_{32}\}$

6.2 実験設定

方策改善性能を評価するための未知環境は図 6.5 に示す．そして，実験パラメータの設定を表 6.14 にまとめる．

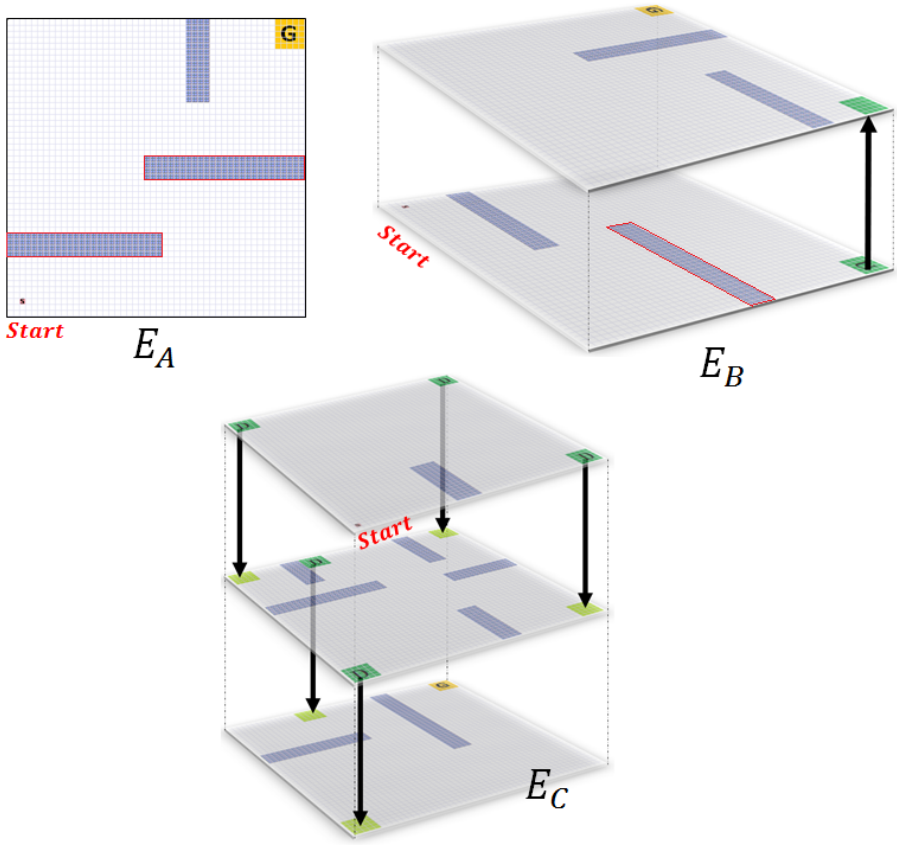


図 6.5: 実験用の環境

表 6.14: パラメータの設定

変数	値	変数	値
γ	0.8	γ_z	0.8
r	non-fix	r_0	100
w_0	10	t	300
m	50	n	15, 25, 35

6.2.1 混合分布の構成要素の設定

$E_1, \ldots, E_{50}$ の環境における 2000 試行の方策学習で得られた同時分布 $P_1, \ldots, P_{50}$ を混合分布の構成要素として，未知環境 E_A, E_B, E_C において利益共有法のみで方策学習を行った場合と，混合分布 (構成要素が m =50 個，$n = 15, 25, 35$ 個) による方策改善を適用した場合で，2000 試行の成功率をもとに評価を行う. なお，$n = 1$ の場合は混合分布とはならないが，同様の手順で方策改善処理を行うことも可能となる.

6.3 結果と考察

未知環境において τ 回の試行で得られる同時分布との Hellinger distance が最小である同時分布が，混合分布の構成要素として選択される. 未知環境 E_A, E_B, E_C において混合分布の構成要素として選択された要素をそれぞれ表 6.15，表 6.16，表 6.17 に示す.

そして，各環境による実験 10 回振り返した中で，2000 試行の平均成功率であった結果の 100 試行毎の成功率の推移をそれぞれ図 6.6 に示す. そして，混合分布 (構成要素が 50 個，$MC = 15, 25, 35$ 個) による方策改善を適用した場合の計算処理時間の結果を表 6.18 に示す.

表 6.15: 環境 E_A において選択された要素

構成要素数	選択された要素
15	$P_{49}, P_5, P_{45}, P_{29}, P_{12}, P_{15}, P_{10}, P_3, P_{18}, P_{26}, P_{35}, P_{48}, P_8, P_{16}, P_{23}$
25	$P_{49}, P_{21}, P_5, P_{13}, P_{45}, P_{29}, P_{12}, P_{15}, P_{25}, P_{19}, P_{10}, P_3, P_7, P_{38}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_8, P_{16}, P_{23}, P_{32}$
35	$P_1, P_{49}, P_{21}, P_5, P_9, P_{13}, P_{45}, P_{29}, P_{12}, P_{20}, P_{15}, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$

表 6.16: 環境 E_B において選択された要素

Z	構成要素数	選択された要素
1	15	$P_{49}, P_{17}, P_{41}, P_{29}, P_{20}, P_{25}, P_{10}, P_3, P_{18}, P_{26}, P_{27}, P_{48}, P_8, P_{16}, P_{32}$
	25	$P_{49}, P_{21}, P_{17}, P_{13}, P_{45}, P_{29}, P_{12}, P_6, P_{25}, P_{19}, P_{10}, P_3, P_7, P_{38}, P_{18}, P_{42}, P_{26}, P_{34}, P_{27}, P_{48}, P_{28}, P_8, P_{16}, P_{23}, P_{32}$
	35	$P_1, P_{49}, P_{21}, P_{17}, P_9, P_{13}, P_{45}, P_{29}, P_{12}, P_{20}, P_6, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{42}, P_{26}, P_{34}, P_{27}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$
2	15	$P_{49}, P_5, P_{41}, P_{29}, P_{12}, P_{25}, P_{10}, P_3, P_{18}, P_{26}, P_{35}, P_{48}, P_8, P_{16}, P_{23}$
	25	$P_{49}, P_{21}, P_5, P_{13}, P_{41}, P_{29}, P_{12}, P_{47}, P_{25}, P_{19}, P_{10}, P_3, P_7, P_{14}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$
	35	$P_1, P_{49}, P_{21}, P_5, P_9, P_{13}, P_{41}, P_{29}, P_{12}, P_{20}, P_{15}, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$

表 6.17: 環境 E_C において選択された要素

Z	構成要素数	選択された要素
1	15	$P_{37}, P_5, P_{45}, P_{29}, P_{12}, P_{33}, P_{10}, P_3, P_{14}, P_{34}, P_{35}, P_{28}, P_8, P_{16}, P_{23}$
	25	$P_{37}, P_{21}, P_5, P_{13}, P_{45}, P_{29}, P_{12}, P_{47}, P_{33}, P_{19}, P_{10}, P_3, P_7, P_{14}, P_{11}, P_{46}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_8, P_{16}, P_{23}, P_{32}$
	35	$P_1, P_{37}, P_{21}, P_5, P_9, P_{13}, P_{45}, P_{29}, P_{12}, P_{20}, P_{15}, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{46}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$
2	15	$P_{37}, P_{13}, P_{41}, P_{29}, P_{12}, P_{25}, P_{10}, P_3, P_{14}, P_{34}, P_{35}, P_{28}, P_8, P_{16}, P_{23}$
	25	$P_{37}, P_{21}, P_5, P_{13}, P_{41}, P_{29}, P_{12}, P_{47}, P_{25}, P_{19}, P_{10}, P_3, P_7, P_{14}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_8, P_{16}, P_{23}, P_{32}$
	35	$P_1, P_{37}, P_{21}, P_5, P_9, P_{13}, P_{41}, P_{29}, P_{12}, P_{20}, P_{15}, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$
3	15	$P_{49}, P_{17}, P_{45}, P_{29}, P_{12}, P_{47}, P_{10}, P_3, P_{11}, P_{34}, P_{35}, P_{28}, P_8, P_{16}, P_{23}$
	25	$P_{49}, P_{21}, P_{17}, P_{13}, P_{45}, P_{29}, P_{12}, P_{47}, P_{33}, P_{19}, P_{10}, P_3, P_7, P_{14}, P_{11}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_8, P_{16}, P_{23}, P_{32}$
	35	$P_1, P_{49}, P_{21}, P_{17}, P_9, P_{13}, P_{45}, P_{29}, P_{12}, P_{20}, P_{24}, P_{47}, P_{25}, P_{33}, P_{19}, P_{10}, P_3, P_{30}, P_7, P_{14}, P_{38}, P_{50}, P_{11}, P_{18}, P_{22}, P_{26}, P_{34}, P_{35}, P_{48}, P_{28}, P_{40}, P_8, P_{16}, P_{23}, P_{32}$

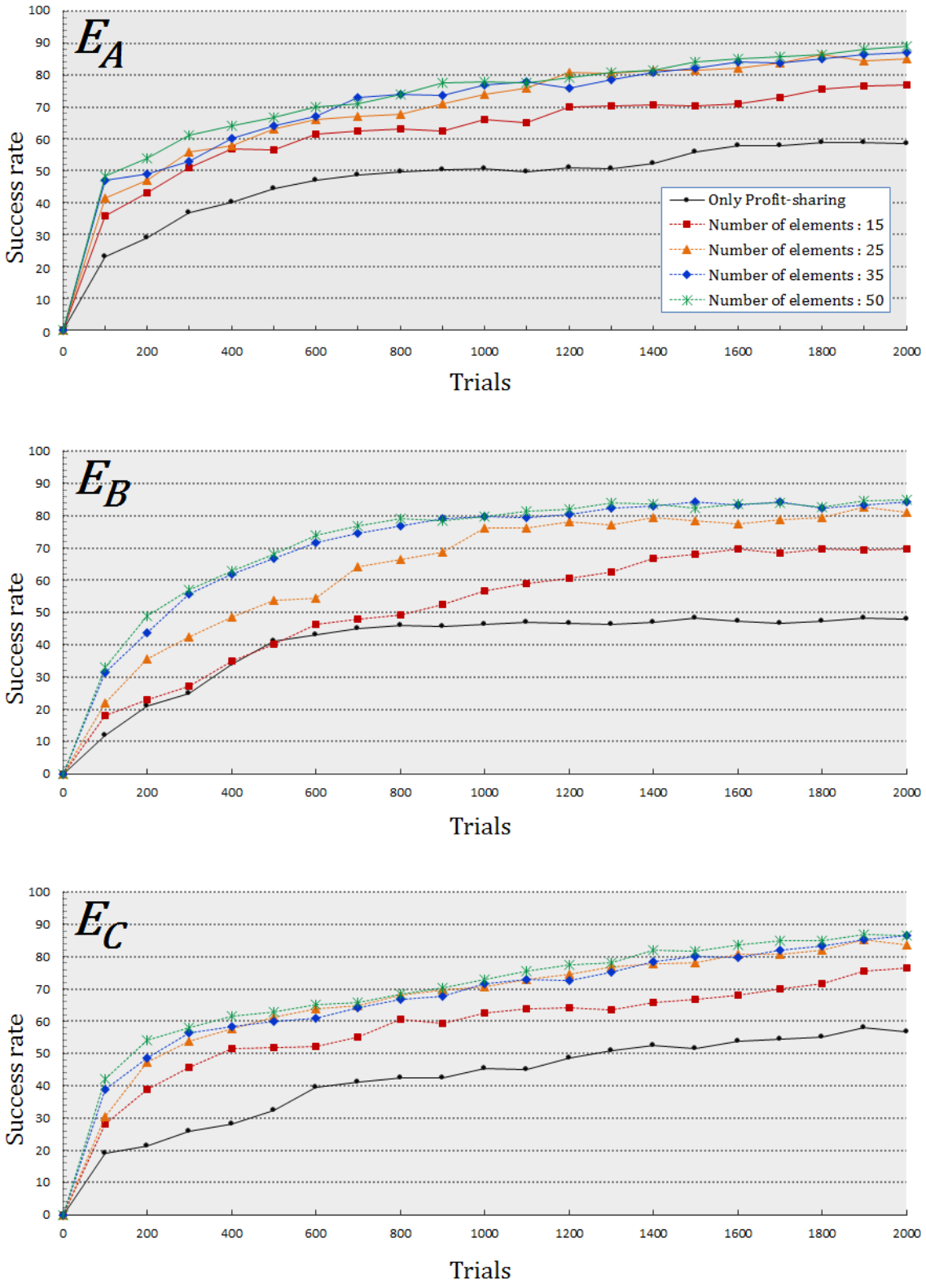


図 6.6: 未知環境における実験 10 回中での平均成功率の推移

表 6.18: 処理時間の比較

実験環境	クラスタ数・処理時間（秒）			
	50	25	25	15
E_A	28.622	23.842	20.445	16.371
E_B	32.319	27.027	23.974	20.113
E_C	37.288	32.124	26.375	23.696

実験結果から， 利益共有法のみによる方策学習より方策改善を適用直後の成功率のほうが明確にいずれも高い．また，方策改善を適用した場合の成功率が実験が終わるまで利益共有法のみの場合よりずっと高く続いていることがわかる．さらに混合分布の構成要素数を減らしても有効性の低下を抑制できることを確認できた．

図 6.6 より、全環境に対して，方策改善を適用した場合の成功率が早い段階 (100 試行) で利益共有法のみによる方策学習の場合の成功率を上回っている．こちらは方策改善を適用した後の未知環境に適応スピードが利益共有法のみより早くなっており，そして，その高い成功率は実験が終了するまで続いており，方策改善の効果が表れていると言える．

環境 E_A と E_C に対して，方策改善を適用した場合の成功率は利益共有法のみによる方策学習の場合の成功率より 20%以上を上回っており，そして，E_B に対しては 30%以上を上回っている．しかし，15 個の構成要素を利用した場合の成功率が利益共有法のみによる方策学習の場合の成功率より高くても，25，35 と 50 個の構成要素を利用した場合の成功率と比較するとまだ低いことが分かった．これより，混合分布の構成要素の数を減少しすぎると，全環境に対しての方策改善への影響が表れた．しかしながら，50 個の構成要素を利用した場合は最も高い成功率が得られても，半分である 25 個までの構成要素を利用した場合でも 50 個を利用した場合とほぼ同じくらいの成功率が得られることが確認できた．そのため，全環境に対して構成要素数の減少による方策改善への有効性の低下は，クラスタリングによって抑制されてい

ると考えられる．

そして，クラスタ数の増加に対して，処理時間の増加は表 6.18 のような結果となっているので，効率的に計算処理時間を短縮していることがわかる．この結果より，混合分布の構成要素数を半分に減少することで，計算処理時間を短縮することができ，そして，方策改善への有効性の低下を抑制することもできると言える．

図 6.7 は 25 個の構成要素を利用した場合と利益共有法のみを利用した場合の結果の比較を示す．この結果より，全環境に対して利益共有法のみを利用した場合はうまく学習できていい結果が得られたときもあったが，学習が失敗して成功率がとても低いときもあった．また，5 回中の成功率の差が非常に大きいので，とても安定性が低いと分かった．一方，混合モデルを利用した場合は成功率が高く，また，5 回中の成功率の差がほとんどないので，学習の安定性が高いと確認できた．

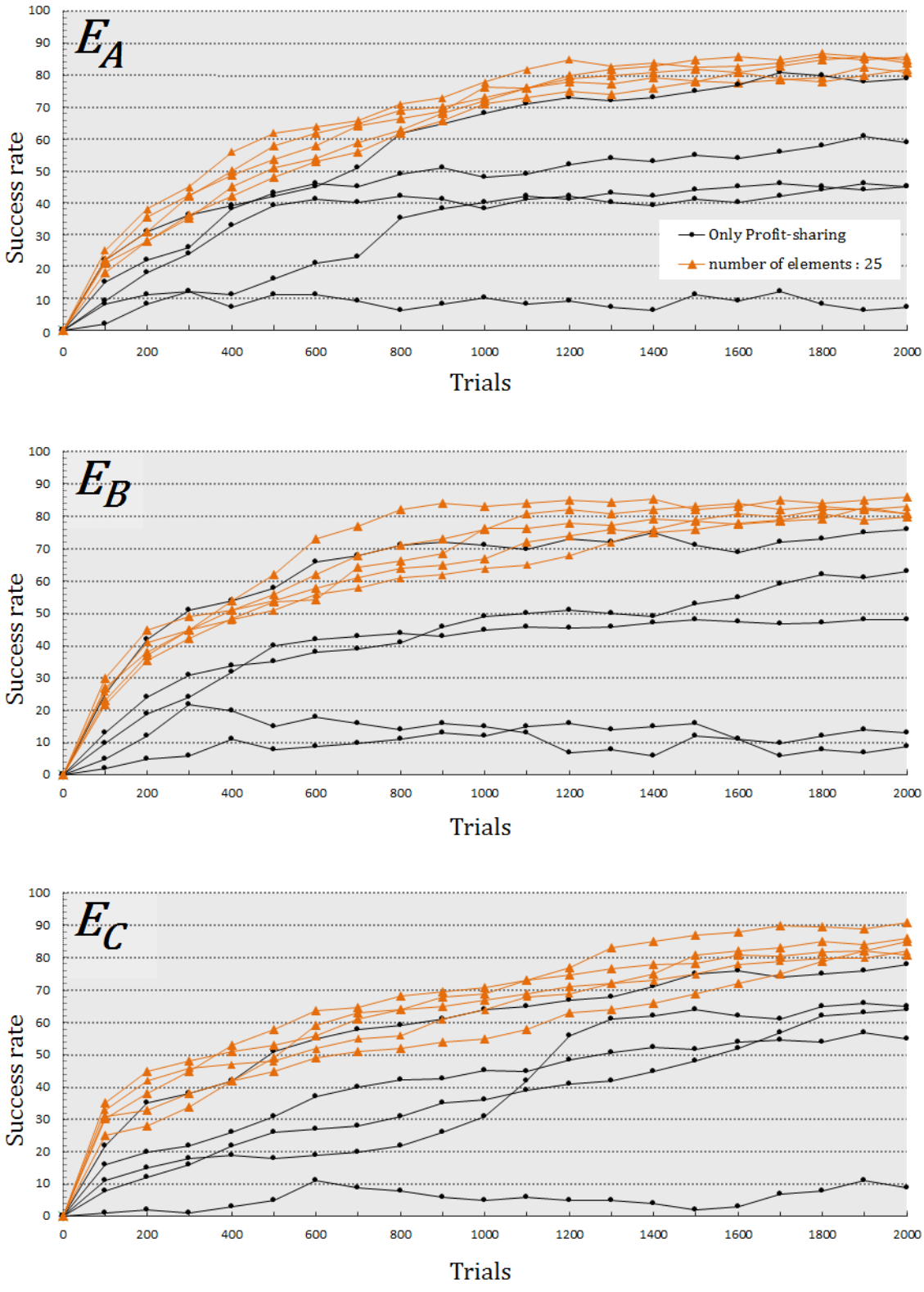


図 6.7: 25 個の構成要素により未知環境における実験 5 回中での成功率の推移

環境 E_C に対するエージェントが目的地に達したまで通った軌跡の試行 1 回〜500 回，501 回〜1000 回および 1001 回〜2000 回の各場合をそれぞれ図 6.8, 図 6.9 と図 6.10 に示す．色の強度 (明るい赤色から濃い赤色まで) は，エージェントが通った軌跡の頻度を示す．そして，各層の階段を通った割合と平均成功率は表 6.19 で表す．

試行 1 回〜500 回の結果（図 6.8）より，エージェントがまだ十分学習されていないため，上の層にエージェントが単純に最も簡単な D-1 に辿りついた．真ん中の層にも同じ様にエージェントが簡単な D-4 に 65%の割合で辿りついた．しかし，D-4 から出発したエージェントは目的地に辿りつくのにかなり難しいため，成功率も 50%までしか上がらなかった．

試行 501 回〜1000 回の結果（図 6.9）より，エージェントが十分学習されいていたため，上の層にエージェントが D-1 より，もっと難しい D-2 と 3 に多く辿りつくようになった．そして，試行 1 回〜500 回の場合と比べて，今回の結果，D-2 と D-3 から出発したエージェントが D-5 に多く辿りつく割合が多いため，成功率が 66%まで上がった．

試行 1001 回〜2000 回の結果（図 6.10）より，D-5 に辿りつくのに，D-2 と D-3 から出発した方が簡単なので，上の層ではエージェントが D-1 に辿りつかずに，ほっとん D-2 と D-3 に辿りつくようになった．また，エージェントが真ん中の層の D-5 に辿りつく割合が 85%までなったので，最終的の成功率も 81%まで上がった．

これらの結果から，非固定の報酬 (r) と偽りの報酬（pseudo-reward）を利用することによって，エージェントが目的に辿りつくように適切な行動を選択することができたので，このような階層型環境にもうまく対応できることを確認できた．

表 6.19: 各階段を通った割合

試行回数	各層の階段を通った割合と平均成功率	
1 回〜500 回	上の層	D-1:46%, D-2:35%, D-3:19%
	真ん中の層	D-4:65%, D-5:35%
	平均成功率: 50%	
501 回〜1000 回	上の層	D-1:22%, D-2:51%, D-3:27%
	真ん中の層	D-4:30%, D-5:70%
	平均成功率: 66%	
1001 回〜2000 回	上の層	D-1:03%, D-2:52%, D-3:45%
	真ん中の層	D-4:15%, D-5:85%
	平均成功率: 81%	

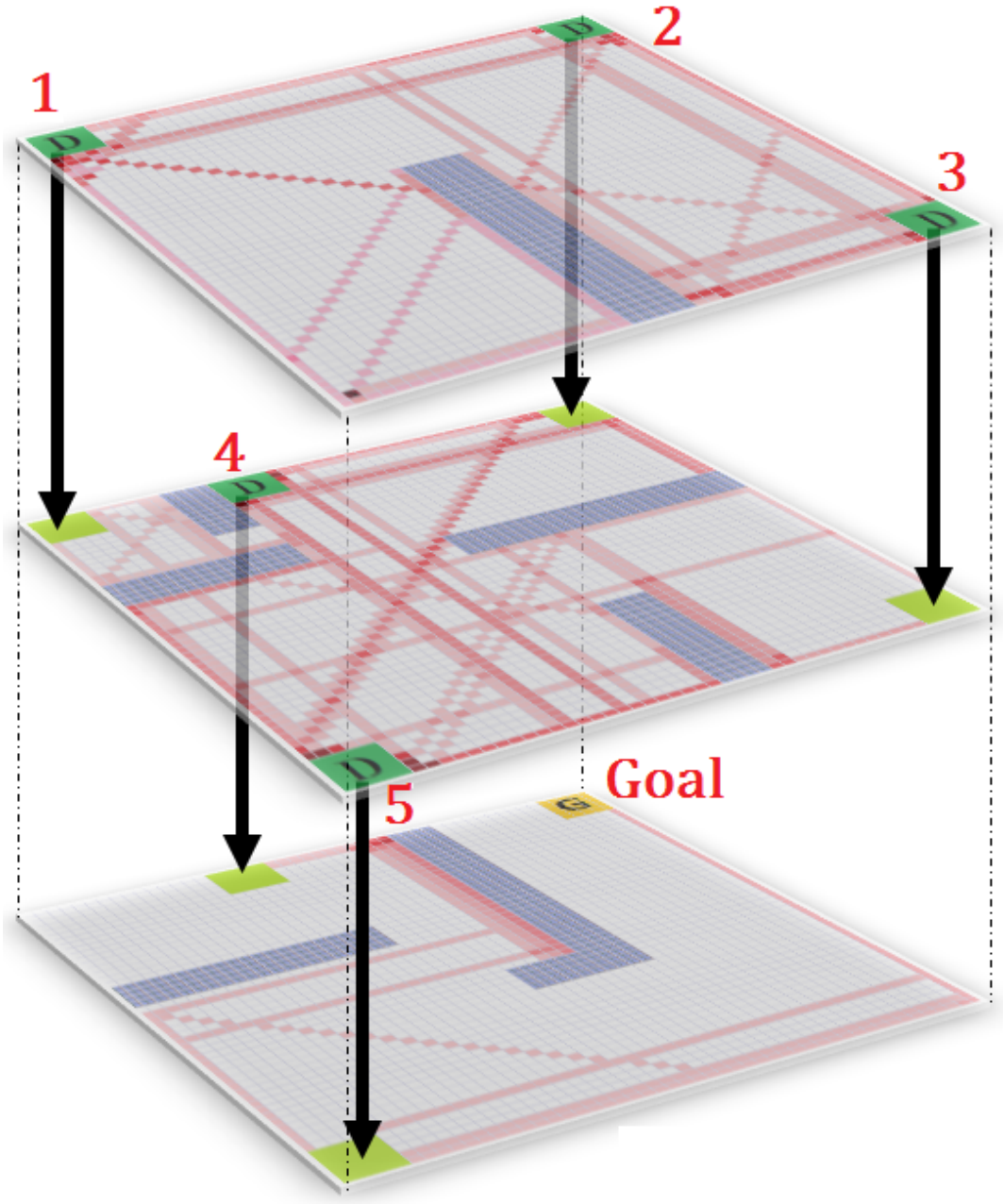


図 6.8: E_C における試行 500 回までのルート

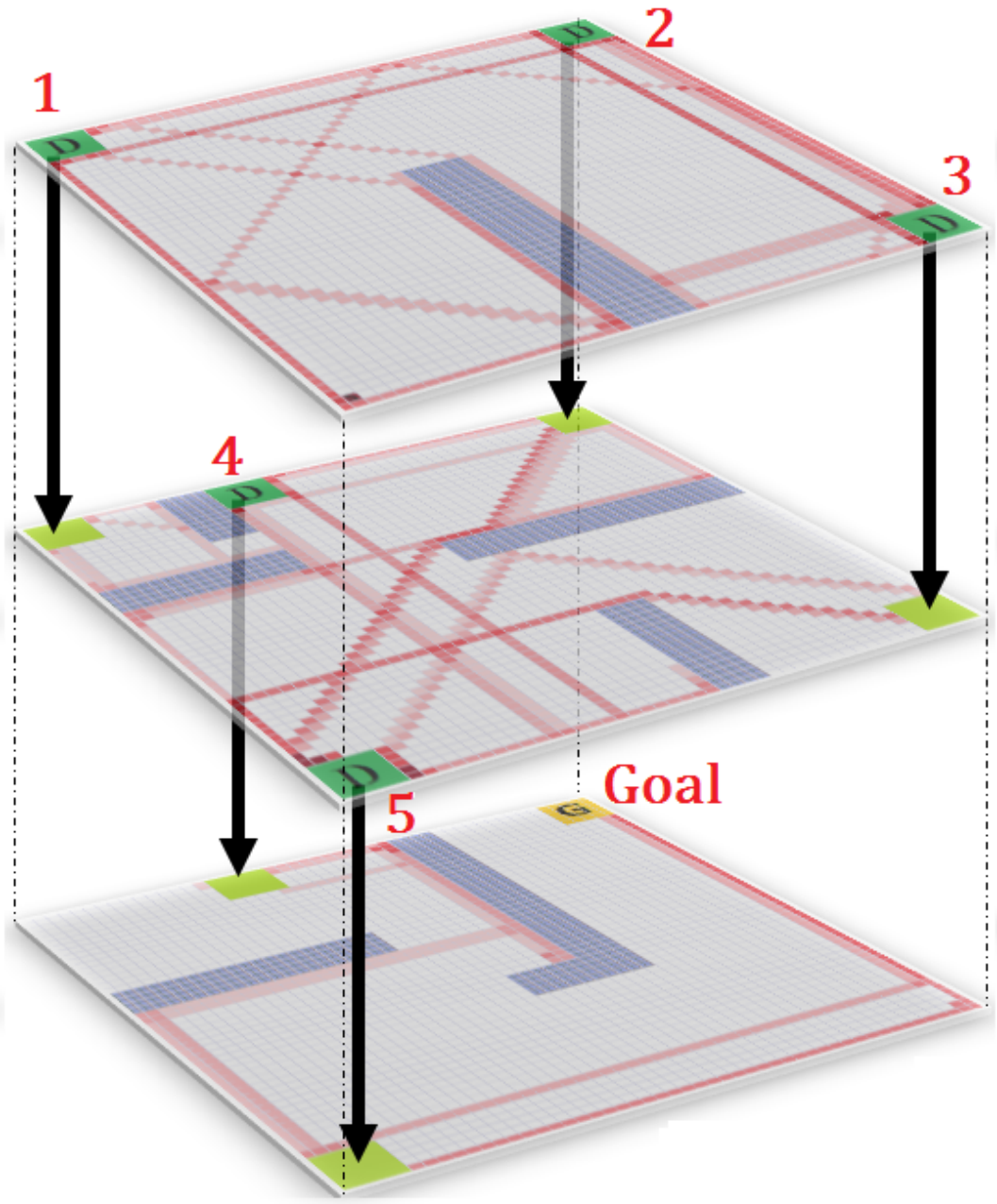


図 6.9: E_C における試行 501 1000 回までのルート

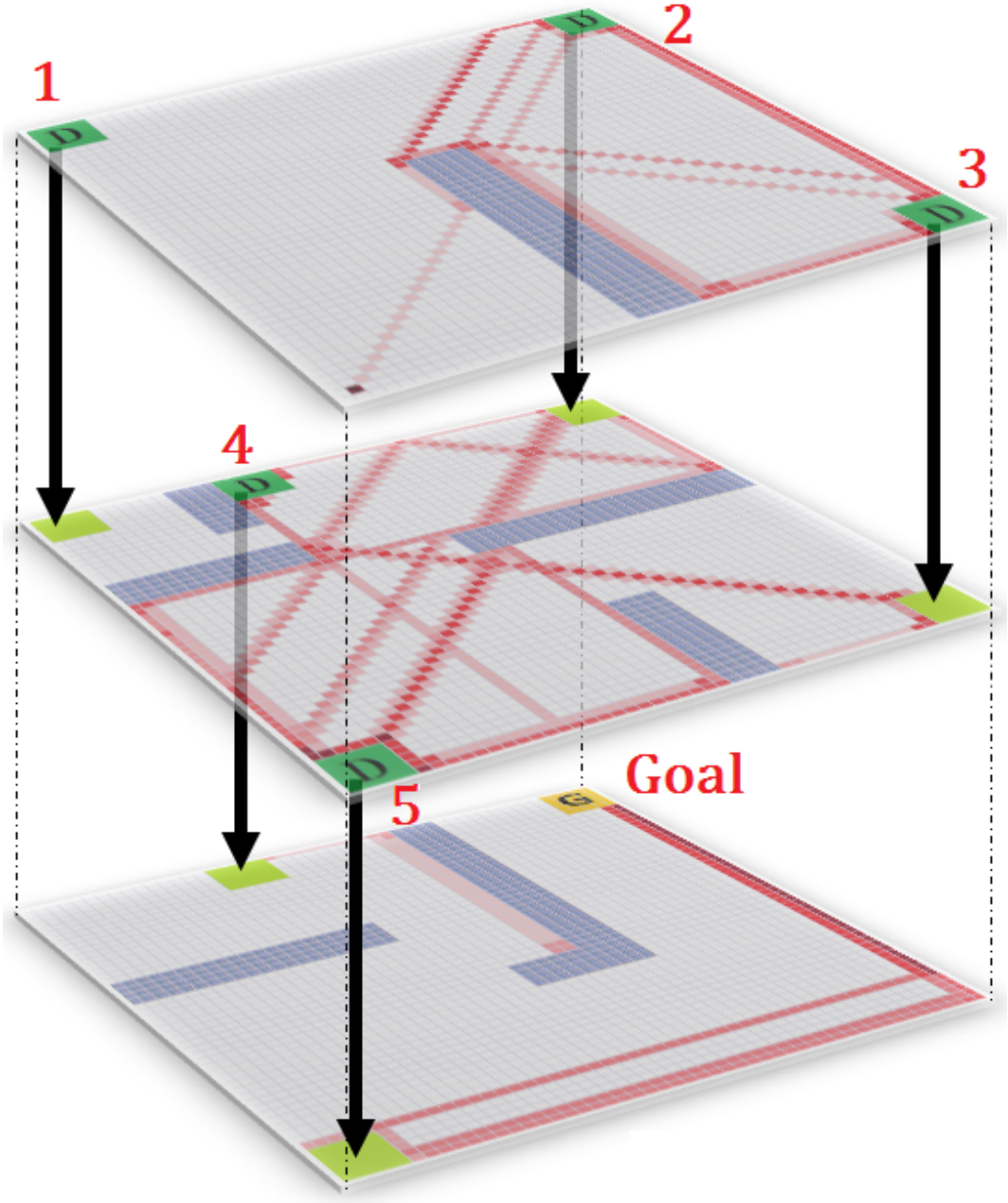


図 6.10: E_C における試行 1001 2000 回までのルート

6.4 追加実験

6.4.1 動的障害物について

動的な障害物を固定，周期的移動，非周期的移動の 3 つの場合を比較するため，本実験では環境 E_A と E_B における 3 つの場合の障害物を適用し，25 個の混合分布による方策改善を行った．変化したパラメータだけを表 6.20 に示す.

表 6.20: 変化パラメータ

変化パラメータ	値		新しい値
n	15, 25, 35	→	25(only)
動的な障害物	周期	→	固定 周期 非周期

実験 10 回振り返した中で，2000 試行の平均成功率であった結果の成功率の推移をそれぞれ図 6.11 に示す．そして，周期的と非周期的の場合のみを 5 回振り返した中で成功率の推移をそれぞれ図 6.12 に示す.

図 6.11 から，両環境 E_A と E_B において障害物が周期的移動する場合の成功率が固定する場合の成功率とほぼ差が無かった．一方，非周期的に移動した場合については，環境 E_B において他の場合の成功率との差がなく，図 6.12 のように，逆に周期的な場合の成功率より上回る結果もあった．しかし，図 6.11 の環境 E_A の場合の結果は動的な環境の場合の成功率よりかなり低かったし，また，図 6.12 のように，全環境において 5 回中の成功率はお互いの差が大きいので安定性な結果が得られなかった.

これらの結果から，エージェントが動的な環境にも適応でき，そして，障害物が非周期に移動した場合より周期的に移動し場合の方が安定性が高くうまく学習できたと分かった．エージェントが動的な障害物より固定な障害物のほうがうまく学習できる場合だけでなく，障害物の位置と目的地の位置など設定によって動的な環境でも障害物が固定した環境よりうまく学習できる場合もあると考えられる.

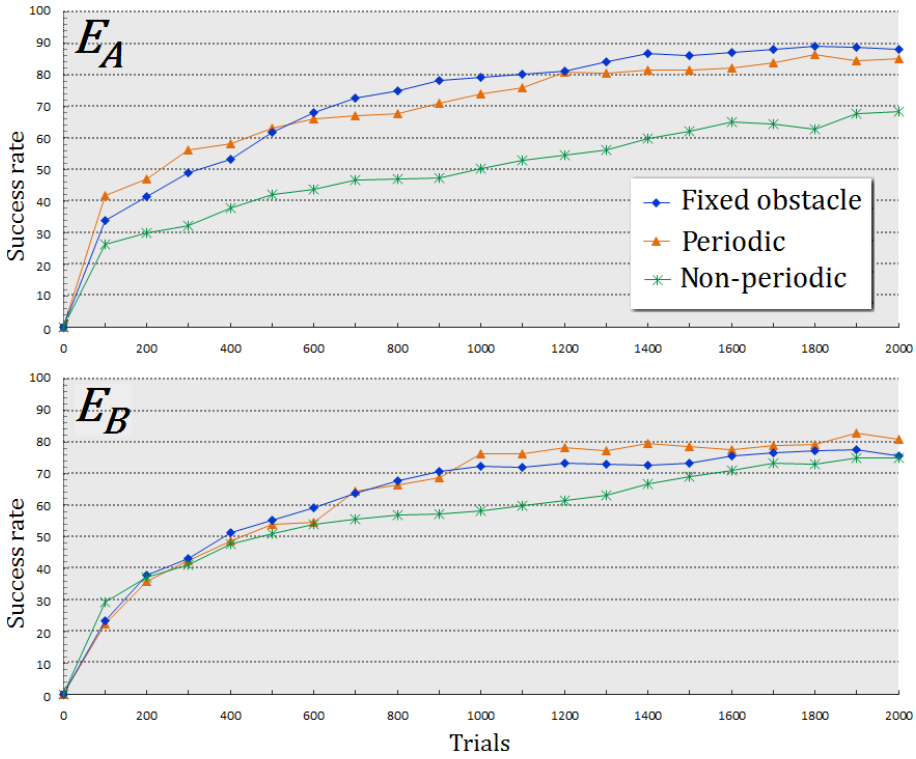


図 6.11: 各方法による要素を選択するイメージ

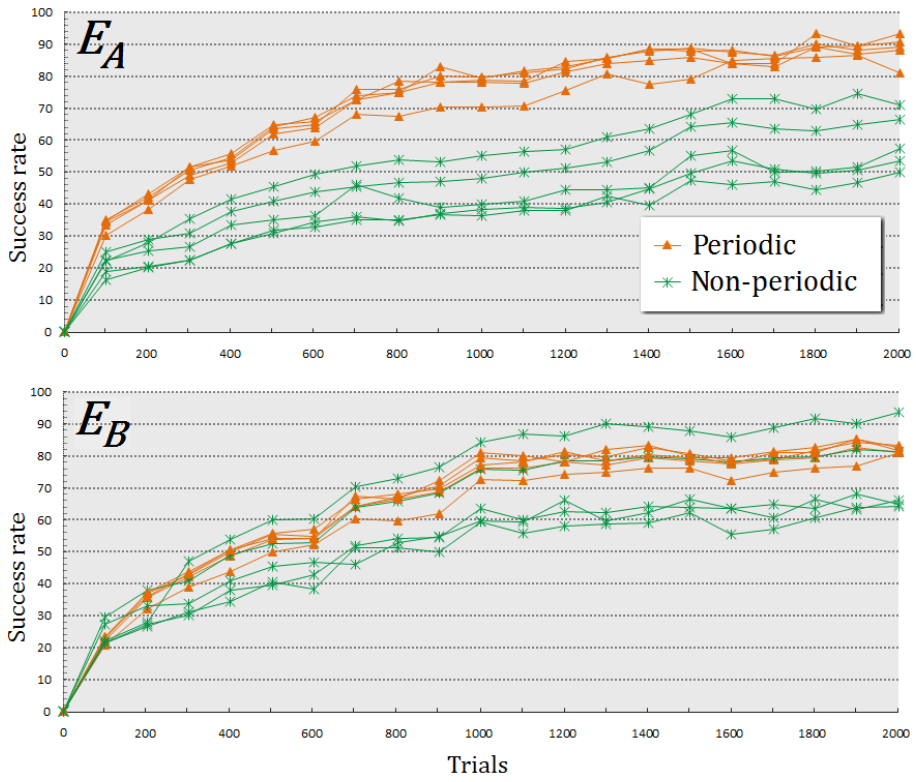


図 6.12: 各方法による要素を選択するイメージ

6.4.2　構成要素の選択について

混合分布の構成要素を選択するのにクラスタリングの方法以外にの方法と比較するため，本実験では未知環境に対して方角学習において，クラスタリングによって構成要素を選択する場合，未知環境で適当の回数の施行で得られたサンプル分布 Q との距離が最も近い順で選択する場合と，ランダムで選択する場合の混合分布による方策改善を行った．

実験の設定

混合分布の構成要素に利用される 12 個の環境と，各場合の方策改善性能を評価するための未知環境を図 6.13 と図 6.14 にそれぞれ示す．そして，実験パラメータの設定を表 6.21 にまとめる．

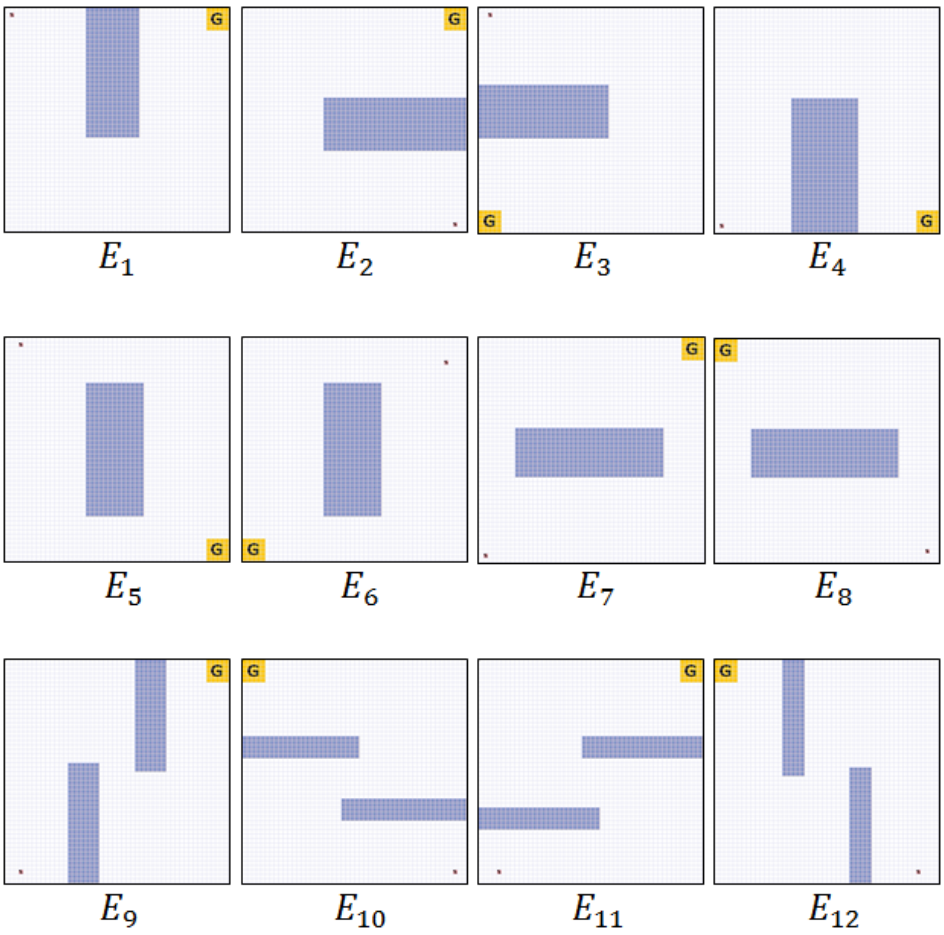


図 6.13: 既知環境

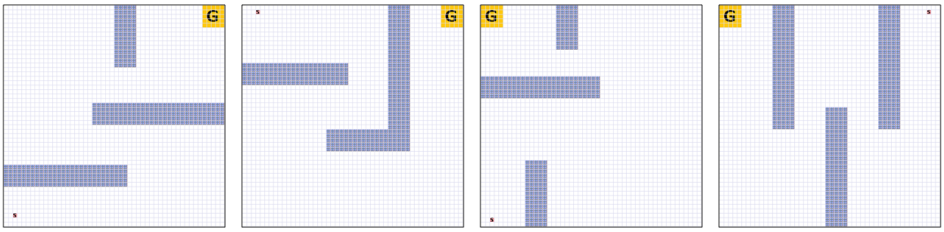


図 6.14: 未知環境

表 6.21: パラメータの設定

変数	値	変数	値
γ	0.8	γ_z	無し
r	non-fix	r_0	100
w_0	10	t	300
m	12	n	6

結果と考察

未知環境 ①, ②, ③, ④ の場合で混合分布の構成要素として選択された要素をそれぞれ表 6.22，表 6.23，表 6.24，表 6.25 に示し，各環境による実験 10 回振り返した中で，2000 試行の平均成功率であった結果の 100 試行毎の成功率の推移を図 6.15 に示す.

結果より，分布クラスタリングによる要素を選択する場合が他の方法を利用する場合よりいい結果が得られたが，環境 ②, ③, ④ の場合の結果は，距離の短い順またはランダムで選択する場合の最終的な成功率の差をクラスタリングの場合と比較するとほとんど無いときもあることがわかる．ただし，一番重要なところでは，どんな未知環境にとっても分布クラスタリングによる要素を選択する場合の成功率が他の方法より早い段階で上がっていくことが確認できた.

環境 ① と ③ の結果から，図 6.16（サンプル分布との距離による選択されたケース）のように，サンプル分布 Q との距離が最も近い順で選択された要素がお互いに似ていると考えられるので，距離の短い順

で選択された要素を利用した場合よりクラスタリングを利用した場合の成功率はかなり高い成功率が得られた．図 6.16（クラスタリングによる選択されたケース 2）のように，クラスタリングによる選択された場合は要素の多様性を維持するために似ている要素を避けることができたので，似ている要素を何回も選択した場合よりいい結果が得られたのは当然である．そして，図 6.16（クラスタリングによる選択されたケース 1）のように，両方法で選択された要素がほぼ同じな場合は環境 ② と ④ の結果のように，両方法での成功率の差もほとんど無かった．

しかしながら，方策改善の性能はたくさんの既知環境から得られる混合分布の構成要素の多様性と未知環境から得られるサンプル分布 Q に依存する．はずれな値を持つ要素がたくさいあると，クラスタリングがうまく機能できない場合があるし，また，構成要素とサンプル分布の距離が大きすぎると，たくさんの構成要素を利用しても，利益共有法のみを利用した場合とほぼ変わらない結果になり，方策改善が機能していない可能性もある [24].

表 6.22: 環境 ① の場合で選択された要素

選択方法	選択された要素
クラスタリング	$P_9, P_{10}, P_3, P_4, P_7, P_{12}$
距離	$P_9, P_{11}, P_5, P_4, P_{10}, P_{12}$
ランダム	$P_1, P_4, P_5, P_6, P_8, P_{12}$

表 6.23: 環境 ② の場合で選択された要素

選択方法	選択された要素
クラスタリング	$P_9, P_{10}, P_3, P_4, P_7, P_{12}$
距離	$P_9, P_{11}, P_5, P_7, P_{10}, P_{12}$
ランダム	$P_3, P_4, P_5, P_7, P_10, P_{12}$

表 6.24: 環境 ③ の場合で選択された要素

選択方法	選択された要素
クラスタリング	$P_6, P_9, P_8, P_4, P_7, P_{12}$
距離	$P_3, P_4, P_7, P_9, P_{10}, P_{12}$
ランダム	$P_1, P_3, P_4, P_8, P_9, P_{11}$

表 6.25: 環境 ④ の場合で選択された要素

選択方法	選択された要素
クラスタリング	$P_{11}, P_6, P_8, P_4, P_7, P_{12}$
距離	$P_{11}, P_4, P_5, P_7, P_{10}, P_{12}$
ランダム	$P_2, P_4, P_5, P_8, P_{10}, P_{11}$

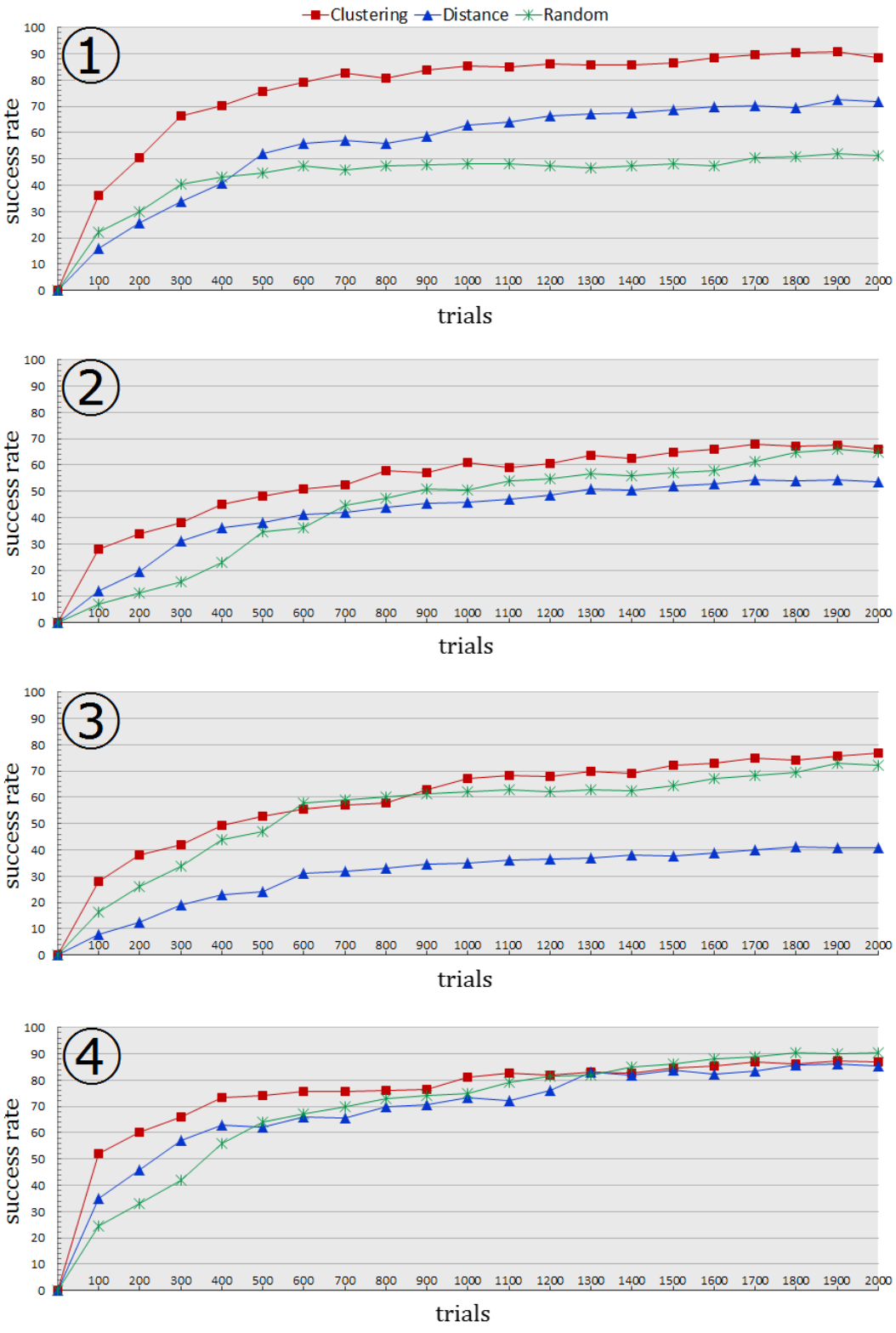
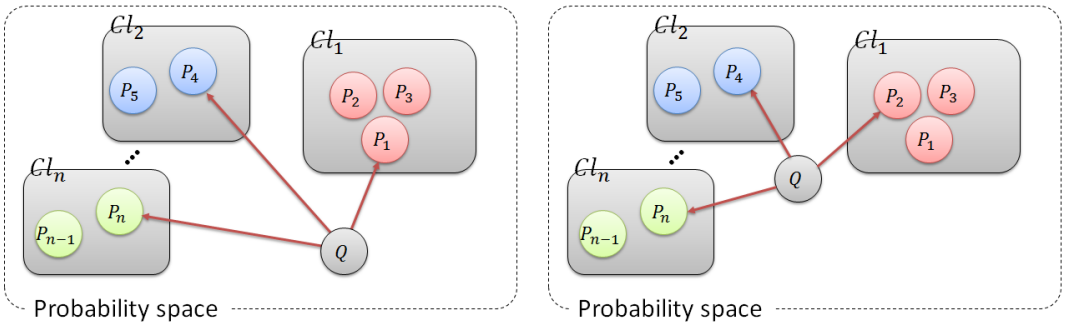
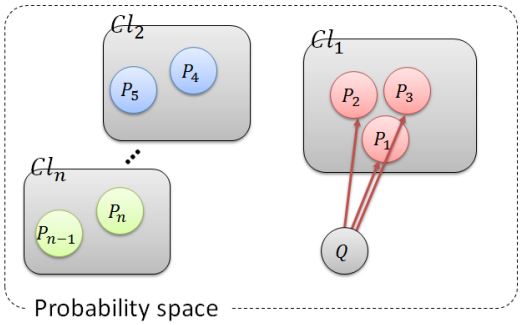


図 6.15: 実験 10 回中での平均成功率の推移



クラスタリングによる選択されたケース1と2



サンプル分布との距離による選択されたケース

図 6.16: 各方法による要素を選択するイメージ

第7章　結論

7.1　まとめ

本研究では，未知環境から得られるサンプル分布Qとの類似度を求めるために，強化学習エージェントが過去の環境で観測したデータからなる同時分布$P(z,s,a)$を利用する．この同時分布$P(z,s,a)$から作られる知識は エージェントが環境の変化での方策を学習するのに非常に有効なものである．サンプル分布Qを獲得するために少し手間がかかるが，エージェントが未知環境において効率的に方策を学習できれば，まだ十分有利だと考えられる．

さらに，クラスタリングを利用することで，似てる分布を避けることができるし，また，構成要素の数を減らしても多様性を維持するこもできるので，適切な構成要素を混合することができた．

提案手法について，エージェントナビゲーション問題を適用例として計算機実験を行い，以下のことを確認した．

- 混合モデルを利用することで，動的な未知環境における方策改善が有効である．
- エピソードを二次元化することで，エージェントが階層環境において適切なルールを選択できた．
- クラスタリングによる選択される混合分布を用いる方法が最も有効である．
- クラスタリングによって，混合分布の構成要素数の減少による方策改善への有効性の低下を抑制可能である．

よって，提案手法より効率的な強化学習システムが構築可能であるといえる．実環境においては，エージェントが学習を試行できる回数

には厳しい制限が生じることとも考えられ，提案手法による学習の安定性と速度の向上，計算量の抑制が有効である．

7.2 今後の課題

今後の課題としては以下のものが挙げられる．

- 混合モデルの構成要素の改良，混合方法の改良．

  混合分布の構成要素に用いた既知環境は2次元の環境しか対応していないため，3次元の環境や他の種類の環境なども利用できるように要素を改良する．また，限られた既知環境の中，似ていない環境にも反対知識として活かせるよう混合方法の改良も行いと思っている．

- ロボット実験．

  本提案手法は動的な階層環境に適応可能な強化学習システムであることを検証できたので，実際にロボットを利用して実験を行いたいと思っている．

謝辞

本研究を進めるにあたり，熱心な御指導と御助言をしていただいた室蘭工業大学大学院工学研究科しくみ情報系領域 塩谷 浩之 教授，東京工業高等専門学校情報工学科 北越 大輔 准教授に心よりお礼申し上げます.

本論文を審査して頂きました室蘭工業大学大学院工学研究科しくみ情報系領域前田 純治 教授，板倉　賢一　教授に厚くお礼申し上げます.

現在までに奨学金を支えてくださる日本政府（文部科学省），公益財団法人佐藤陽国際奨学財団，公益財団法人平和中島財団，財団法人本庄国際奨学財団の皆様方に、多大なる感謝の意を評したいと思っております.

自分の家族のようにいつも面倒を見てくださっている今浦家，学業のことでもプライベートのことでもいつもサポートしてくださっている成田さん，塩崎さん，そして国際交流センターの皆さん，国際交流クラブの皆さんにも心よりお礼申し上げます.

研究のみならず，様々の面でお世話になったり，一緒に研究室に配属され苦楽を共にした逢坂 光司君，加藤 甫君，川田 結衣さん，西田 佳史君，そして後輩たちにも心より深く感謝致します.

平成 26 年 11 月
ポッマサク ウタイ

参考文献

[1] D.Kitakoshi，H.Shioya and R.Nakano，”Empirical analysis of an on-line adaptive system using a mixture of Bayesian networks”，Information Science，Vol.180，15，pp．2856-2874，2010.

[2] 北越 大輔，塩谷 浩之，中野 良平，”BN 混合モデルを用いたオンライン型方策改善システムの動的環境への適応”，信学技報，Vol.104，No.249，pp. 15-20，2004.

[3] 山口　晃昌 良平，北越 大輔，塩谷 浩之，”方策改善システムにおける環境に関する確率的な知識の視覚化”，平成 18 年度電気・信通信関係学会北海道支部連合大会講演論文集，pp．194，2006.

[4] Dominik M. Endres，and Johannese E. Schindelin，”A New Metric for Probability Distribution”，IEEE Trans. Inform. Theory，Vol.49，No.7，pp．1858-1860，2003.

[5] K. Miyazaki, T. Terada and H. Kobayashi, ”Generating Cooperative Behavior by Multi-Agent Profit Sharing on the Soccer Game”, Proc. cf 4th Int. Symp. On Adva- nced AI Systems, pp. 166-169, 2003.

[6] J.J. Grefenstette, ”Credit assignment in rule discovery systems based on genetic algorithms”, Machine Learning 3 (1988)225-245.

[7] R.S. Sutton ”Temporal Credut Assignment in Reinforcement Learning”, University of Massachusetts, Amherst, MA, 1984.

[8] R.S. Sutton, A.G. Barto, ”Reinforcement Learning: An Introduction”, MIR Press, Cambridge, MA, 1998.

[9] Peters Jan, Sethu Vijayakumar, Stefan Schaal, ”Reinforcement Learning for Humanoid Robotics”, IEEE-RAS International Conference on Humanoid Robots, 2003

[10] C. J. C. H. Watkins, P. Dayan, ”Technical Note: Q-Learning”, Machine Learning 8, pp. 279–292 (1992).

[11] 宮崎和光，木村 元，小林重信，”Profit Sharing に基づく強化学習の理論と応用”, 人工知能学会誌, Vol.14, No.5, pp.800–807 (1999).

[12] R. Parr, S. Russell, "Reinforcement Learning with Hierarchies of Machines", Advances in Neural Information Processing Systems 10, pp. 1043–1049 (1998).

[13] Frühwirth-Schnatter, "Finite Mixture and Markov Switching Models", Springer, ISBN 978-1-4419-2194-9, 2006.

[14] E. Hellinger, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen", Journal of Reine Angewandte Mathematics 136 (1909) pp. 210-271.

[15] M.S. Nikulin,"Hellinger distance", in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4, 2001

[16] K. Bailey, "Numerical Taxonomy and Cluster Analysis", Typologies and Taxonomies, p. 34, ISBN 9780803952591, 1994

[17] Estivill-Castro, Vladimir, "Why so many clustering algorithms — A Position Paper", 2002

[18] T. Hastie, R. Tibshirani, J. Friedman, "14.3.12 Hierarchical clustering", The Elements of Statistical Learning (2nd ed.), New York: Springer. pp. 520—528. ISBN 0-387-84857-6, 2009.

[19] T. Croonenborghs, J. Ramon, H. Blockeel, M. Bruynooghe, "Model-assisted approaches for relational reinforcement learning: some chalenges for the SRL community", Proc. of the ICML-2006 Workshop on Open Problems in Statistical Relational Learning, Pittsburgh, PA, 2006.

[20] F. Fernández, M. Veloso, "Probabilistic policy reuse in a reinforcement learning agent", Proc. of the fifth Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems, 2006, pp. 720-727.

[21] 北越大輔，山口晃昌，塩谷浩之， "クラスタリングを用いた強化学習システム IPMBN の環境変化への適応について"，電子情報通信学会技術研究報告， NC2006-99， pp. 65-70, 2007.

[22] PHOMMASAK Uthai，北越 大輔，塩谷浩之，"強化学習エージェントにおける分布クラスタリングを用いた方策改善に関する検討"， 電気・情報関係学会北海道支部連合大会講演論文集，(CD-ROM)，2011.

[23] Uthai Phommasaki，Daisuke Kitakoshi，Hiroyuki Shioya，"An Adaptation System to Unknown Environment by Modifying the Parameters of the Profit-Sharing Method and Mixture Probability"，IWACIII2011，(CD-ROM)，Paper No. IWACIII-041，2011.

[24] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, “ An adaptation System in Unknown Environments Using a Mixture Probability Model and Clustering Distributions ”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.6, pp. 733-740, 2012.

[25] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, “ A policy-improving system with a mixture probability and clustering distributions to unknown 3d-environments ”, Computer Science and Engineering Conference (ICSEC), pp. 381-386, 2013.

[26] Uthai Phommasak, Daisuke Kitakoshi, Jun Mao, Hiroyuki Shioya, “ A Policy-Improving System for Adaptability to Dynamic Environments Using Mixture Probability and Clustering Distribution ”, Journal of Computer and Communications, pp. 210-219, 2014

[27] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, Junji Maeda, “ A Reinforcement Learning System to Dynamic Movement and Multi-Layer Environments”, Journal of Intelligent Learning Systems and Applications, pp. 210-219, 2014

学会発表歴

学会発表歴 1

学会　：　平成 23 年度電気・情報関係学会北海道支部連合大会
発表日　：　2011 年 10 月 23 日
開催地　：　公立はこだて未来大学
題目　：　強化学習エージェントにおける分布クラスタリングを用いた方策改善に関する検討
著者　：　Uthai PHOMMASAK，北越 大輔，塩谷 浩之
掲載誌　：　平成 23 年度電気・情報関係学会北海道支部連合大会講演論文集 (CD-ROM) 講演番号 196.

学会発表歴 2

学会　：　International Workshop on Advanced Computational Intelligence and IntelligentInformatics 2011 (IWACIII 2011)
発表日　：　21^{th} August 2011
開催地　：　Headquarters of Suzhou University，Suzhou，CHINA
題目　：　An adaptation system to unknown environment by modifying parameters of the profit-sharing method and mixture probability
著者　：　Uthai PHOMMASAK，Daitsuku KITAKOSHI，Hiroyuki SHIOYA
掲載誌　：　International Workshop on Advanced Computational Intelligence and IntelligentInformatics 2011，GS2-1，(CD-ROM) Paper No．IWACIII-041

学会発表歴 3

学会　：2013 International Computer Science and Engineering Conference (ICSEC2013)
発表日　：6^{th} September 2013
開催地　：Silpakorn University, Bangkok, THAILAND
題目　：A Policy-Improving System with Mixture Probability and Clustering Distributions to Unknown 3D-environmetns
著者　：Uthai PHOMMASAK，Daitsuku KITAKOSHI，Hiroyuki SHIOYA
掲載誌　：2013 International Computer Science and Engineering Conference, 2013, pp.381-386

学会発表歴 4

学会　：The 2^{nd} Conference on Artificial Intelligence and Data Mining (AIDM 2014)
発表日　：11^{th} March 2014
開催地　：Youngor central Hotel，Suzhou，CHINA
題目　：A Policy-Improving System to Dynamic Environments by using Mixture Probability andClustering Distributions
著者　：Uthai PHOMMASAK，Daitsuku KITAKOSHI，
Jun Mao, Hiroyuki SHIOYA
掲載誌　：The 2^{nd} Conference on Artificial Intelligence and Data Mining, 2014, pp.210-219

研究業績

[**1**] 栞田 充邦, 沢田 和也, 松本 優幸, ポッマサク ウタイ, ヤスィン エルバダウィ, 長尾 和彦, “ネットワーク監視システムにおける監視情報の“見える化”に関する考察”, 電子情報通信学会総合大会講演論文集 2008 年_通信 (2), “S-118” “S-119”, 2008.

[**2**] Uthai Phommasak, 北越 大輔, 塩谷 浩之, “強化学習エージェントにおける分布クラスタリングを用いた方策改善に関する検討”, 電気・情報関係学会北海道支部連合大会講演論文集 (CD-ROM), 巻：2011 ページ：ROMBUNNO.196, 2011.

[**3**] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, “An adaptation System in Unknown Environments by modifying parameters of Profit-sharing method and Mixture Probability”, International Workshop on Advanced Computational Intelligence and Intelligent Informatics IWAIII2011, GS2-1, (CD-ROM) Paper No.IWACIII-041.

[**4**] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, “An adaptation System in Unknown Environments Using a Mixture Probability Model and Clustering Distributions”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.6, pp. 733-740, 2012.

[**5**] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, “A policy-improving system with a mixture probability and clustering distributions to unknown 3d-environments”, Computer Science and Engineering Conference (ICSEC), pp. 381-386, 2013.

[**6**] Uthai Phommasak, Daisuke Kitakoshi, Jun Mao, Hiroyuki Shioya, “A Policy-Improving System for Adaptability to Dynamic Environments Using Mixture Probability and Clustering Distribution”, Journal of Computer and Communications, pp. 210-219, 2014

[**7**] Uthai Phommasak, Daisuke Kitakoshi, Hiroyuki Shioya, Junji Maeda, “A Reinforcement Learning System to Dynamic Movement and Multi-Layer Environments”, Journal of Intelligent Learning Systems and Applications, pp. 210-219, 2014

[8] Jun Mao, Uthai Phommasak, Shinya Watanabe, Hiroyuki Shioya, “ Detecting Foggy Images and Estimating the Haze Fegree Factor”, Journal of Computer Sciense Systems Biology, Vol.7, No.6, pp. 226-228, 2014