(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4853997号
(P4853997)
(45)発行日　平成24年1月11日(2012.1.11)　(24)登録日　平成23年11月4日(2011.11.4)

(51)Int.Cl.　F I
G06N 3/00 (2006.01)　G06N 3/00 550E

請求項の数 6 (全 23 頁)

(21)出願番号　特願2005-236604(P2005-236604)
(22)出願日　平成17年8月17日(2005.8.17)
(65)公開番号　特開2007-52589(P2007-52589A)
(43)公開日　平成19年3月1日(2007.3.1)
審査請求日　平成20年3月27日(2008.3.27)

（出願人による申告）平成１７年度独立行政法人情報通信研究機構、研究テーマ「人間情報コミュニケーションの研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100067828
弁理士　小谷　悦司
(74)代理人　100096150
弁理士　伊藤　孝夫
(74)代理人　100109438
弁理士　大月　伸介
(72)発明者　杉本　徳和
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　鮫島　和行
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
最終頁に続く

(54)【発明の名称】エージェント学習装置、エージェント学習方法及びエージェント学習プログラム

(57)【特許請求の範囲】
【請求項１】
環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習装置であって、
観測関数を通した観測変数として環境の状態を観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化手段と、
前記環境抽象化手段により抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定手段と、
前記状態決定手段により決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定手段と、
下位報酬として連続関数である複数の下位報酬関数の中から、前記状態決定手段により決定されたインデックスと、前記行動決定手段により決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択手段と、
前記下位報酬選択手段により選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力を用いて環境に対して働きかける制御出力決定手段とを備え、
前記環境抽象化手段は、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、前記状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大き

な値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力し、
前記状態決定手段は、前記責任信号を最大にするインデックスを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとすることを特徴とするエージェント学習装置。
【請求項２】
前記行動決定手段は、行動価値関数の状態を指定するインデックスとして前記責任信号を最大にするインデックスを有し且つ前記責任信号により重み付けされた当該行動価値関数の値を最大とする行動を指定するインデックスを、取るべき一の行動を指定するインデックスとすることを特徴とする請求項１記載のエージェント学習装置。
【請求項３】
前記制御出力決定手段は、下位報酬関数を最大にする際に、前記状態予測モデルと下位報酬関数との組み合わせにより構成される下位価値関数の値を最大にするように制御出力を決定することを特徴とする請求項１または２に記載のエージェント学習装置。
【請求項４】
前記下位報酬関数は、環境の状態と制御出力とによって張られる関数空間において、前記状態決定手段により決定されたインデックスと、前記行動決定手段により決定されたインデックスとにより指定される離散的な領域内に頂点を持つ凸関数であることを特徴とする請求項１乃至３のいずれかに記載のエージェント学習装置。
【請求項５】
環境抽象化手段、状態決定手段、行動決定手段、下位報酬選択手段および制御出力決定手段を備えるエージェント学習装置を用いて、環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習方法であって、
環境の状態を観測関数を通した観測変数として観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化ステップと、
前記環境抽象化ステップにより抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定ステップと、
前記状態決定ステップにより決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定ステップと、
下位報酬として連続関数である複数の下位報酬関数の中から、前記状態決定ステップにより決定されたインデックスと、前記行動決定ステップにより決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択ステップと、
前記下位報酬選択ステップにより選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力を用いて環境に対して働きかける制御出力決定ステップとを含み、
前記環境抽象化ステップでは、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、前記状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大きな値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力し、
前記状態決定ステップでは、前記責任信号を最大にするインデックスを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとすることを特徴とするエージェント学習方法。
【請求項６】
環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習プログラムであって、
観測関数を通した観測変数として環境の状態を観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化手段と、
前記環境抽象化手段により抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定手段と、

前記状態決定手段により決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定手段と、

下位報酬として連続関数である複数の下位報酬関数の中から、前記状態決定手段により決定されたインデックスと、前記行動決定手段により決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択手段と、

前記下位報酬選択手段により選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力を用いて環境に対して働きかける制御出力決定手段としてコンピュータを機能させ、

前記環境抽象化手段は、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、前記状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大きな値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力し、

前記状態決定手段は、前記責任信号を最大にするインデックスを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとすることを特徴とするエージェント学習プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、主に強化学習に用いられ階層構造を有するエージェント学習装置、エージェント学習方法及びエージェント学習プログラムに関するものである。

【背景技術】

【０００２】

強化学習は、エージェントが環境からの報酬信号をもとに、自律的に行動則を獲得できるという優れた特性を有する学習理論である。しかしながら、これまでの強化学習は学習速度が遅いため、対象となる環境が複雑な場合は実行不可能なほど膨大な学習時間を要し、実用的な手法ではないとの批判を受けてきた。このため、強化学習の学習速度を速める手段として階層構造の導入が注目されており、階層構造を用いたいくつかの強化学習手法が提案されている。

【０００３】

最も簡単な階層構造を用いた強化学習装置は、上位層と下位層との二層から構成される。下位層は、上位層と環境との相互作用を取り持つ機能を有し、主に環境を抽象化して上位層にその情報を伝える。そして、上位層は、下位層によって抽象化された状態空間において学習を行う。この階層構造を用いることで、報酬の伝搬の速い上位層は下位層から環境の状態を受け取り強化学習を行う。この構成により、当該強化学習装置は高速に行動則を獲得することができる。

【０００４】

階層構造を用いた強化学習手法として、例えば、解像度の粗い上位層を設けることで効率よく探索を行う手法（非特許文献１参照）や、マルコフ決定過程（Ｍａｒｋｏｖ　Ｄｅｃｉｓｉｏｎ　Ｐｒｏｃｅｓｓｅｓ，ＭＤＰ）を分解し、それぞれに価値関数を割り当てる手法（非特許文献２参照）が提案されている。

【非特許文献１】ピー　ダヤン（P. Dayan）他、「封建的な強化学習」（Feudal reinforcement learning）、ニューラルインフォメーションプロセスシステムの進歩（Advances in Neural Information Processing Systems）、１９９３年

【非特許文献２】ティー　ジー　ディエテリッチ（T. G. Dietterich）、「MAXQ価値関数の分解を用いた階層的な強化学習」（Hierarchical reinforcement learning with the MAXQ value function decomposition）、アーティフィシャルインテリジェンスリサーチジャーナル（Journal of Artificial Intelligence Research）、２０００年、vol.13、p. 227 - p. 303

【発明の開示】

【発明が解決しようとする課題】
【０００５】
しかしながら、上記の従来手法は、離散時間及び離散状態を対象としたものであるが、実世界においてヒトやロボットが環境と相互作用を行う場合に観測される環境の状態や環境からの報酬信号、ロボット等を制御するために出力する制御信号等は、一般に連続的な値を取るため、上記従来の手法をそのまま用いても、実世界における連続時間及び連続状態を取り扱うことができない。
【０００６】
また、実験者が連続的な環境を予め離散化してロボットに入力したりすることにより、階層構造を作り込む手法も考えられるが、この手法では、強化学習において最も困難な課題である「環境の抽象化」を実験者が事前知識に基づいて解決しているため、一般性があるとは言えず、連続的である環境の状態を自律的に学習することはできない。
【０００７】
本発明の目的は、連続時間及び連続状態を取り扱うことができるとともに、環境から自律的且つ高速に学習を行うことができるエージェント学習装置、エージェント学習方法及びエージェント学習プログラムを提供することである。
【課題を解決するための手段】
【０００８】
本発明に係るエージェント学習装置は、環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習装置であって、観測関数を通した観測変数として環境の状態を観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化手段と、環境抽象化手段により抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定手段と、状態決定手段により決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定手段と、下位報酬として連続関数である複数の下位報酬関数の中から、状態決定手段により決定されたインデックスと、行動決定手段により決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択手段と、下位報酬選択手段により選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力でもって環境に対して働きかける制御出力決定手段とを備えるものである。
【０００９】
本発明に係るエージェント学習装置では、環境が連続状態から離散状態へと抽象化され、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスが決定されるとともに、取るべき一の行動を指定するインデックスが決定され、これら２つのインデックスにより指定される一の下位報酬関数が選択され、選択された下位報酬関数を最大にする制御出力を用いて環境に働きかけているので、一般に連続状態を有する環境を離散状態へと抽象化し、当該離散状態において強化学習を行うことができ、高速に学習を行うことができる。また、環境に働きかける際には、その離散状態を再度連続状態へと変換しているので、連続時間及び連続状態を取り扱うことができるとともに、環境から自律的且つ高速に学習を行うことができる。
【００１０】
また、環境抽象化手段は、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大きな値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力する。したがって、複数の状態予測モデルが算出する環境の状態変化の予測値はそれぞれ異なるため、現在の環境を最もよく抽象化している一の状態を有効に決定できる。
【００１１】
また、状態決定手段は、責任信号を最大にするインデックスを、現在時刻において最も

環境をよく抽象化している一の状態を指定するインデックスとする。したがって、現在の環境を最もよく抽象化している一の状態を確実に決定できる。

【００１２】

行動決定手段は、行動価値関数の状態を指定するインデックスとして責任信号を最大にするインデックスを有し且つ責任信号により重み付けされた当該行動価値関数の値を最大とする行動を指定するインデックスを、取るべき一の行動を指定するインデックスとすることが好ましい。この場合、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを有する状態の中で、最適な行動を有効に決定することができる。

【００１３】

制御出力決定手段は、下位報酬関数を最大にする際に、状態予測モデルと下位報酬関数との組み合わせにより構成される下位価値関数の値を最大にするように制御出力を決定することが好ましい。この場合、環境に与えられる制御出力として、下位報酬関数を最大にするようなものが選ばれるので、強化学習を有効に行うことができる。

【００１４】

下位報酬関数は、環境の状態と制御出力とによって張られる関数空間において、状態決定手段により決定されたインデックスと、行動決定手段により決定されたインデックスとにより指定される離散的な領域内に頂点を持つ凸関数であることが好ましい。この場合、関数空間において、隣り合った下位報酬関数同士の区別が明確になるために、抽象化された状態間の遷移が確定的になるため、学習を効率的に行うことができる。

【００１５】

本発明に係るエージェント学習方法は、環境抽象化手段、状態決定手段、行動決定手段、下位報酬選択手段および制御出力決定手段を備えるエージェント学習装置を用いて、環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習方法であって、観測関数を通した観測変数として環境の状態を観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化ステップと、環境抽象化ステップにより抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定ステップと、状態決定ステップにより決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定ステップと、下位報酬として連続関数である複数の下位報酬関数の中から、状態決定ステップにより決定されたインデックスと、行動決定ステップにより決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択ステップと、下位報酬選択ステップにより選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力を用いて環境に対して働きかける制御出力決定ステップとを含み、前記環境抽象化ステップでは、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、前記状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大きな値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力し、前記状態決定ステップでは、前記責任信号を最大にするインデックスを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとするものである。

【００１６】

本発明に係るエージェント学習プログラムは、環境に基づいて学習し且つ当該学習の結果に基づいて当該環境に対して働きかけるエージェント学習プログラムであって、観測関数を通した観測変数として環境の状態を観測し、当該観測変数に基づいて環境を連続状態から離散状態へと抽象化する環境抽象化手段と、環境抽象化手段により抽象化された後の離散状態の中から、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスを決定する状態決定手段と、状態決定手段により決定された一の離散状態において学習を行うことで取るべき一の行動を指定するインデックスを決定する行動決定手段と、下位報酬として連続関数である複数の下位報酬関数の中から、状態決定手段により

決定されたインデックスと、行動決定手段により決定されたインデックスとを有する一の下位報酬関数を選択する下位報酬選択手段と、下位報酬選択手段により選択された下位報酬関数を最大にするように環境への制御出力を決定し、当該制御出力を用いて環境に対して働きかける制御出力決定手段としてコンピュータを機能させ、前記環境抽象化手段は、現在時刻における環境の状態と環境への制御出力とに基づいてモデルごとに異なる環境の状態変化の予測値を算出する複数の状態予測モデルにより環境を抽象化し、前記状態予測モデルにより算出された環境の状態変化の予測値と観測変数から推定された環境の状態変化の推定値との差に基づく状態予測誤差が小さいものほど大きな値を取る責任信号を算出し、当該責任信号を抽象化された環境ごとに出力し、前記状態決定手段は、前記責任信号を最大にするインデックスを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとするものである。

【発明の効果】

【００１７】

本発明によれば、一般に連続状態を有する環境を離散状態へと抽象化し、当該離散状態において強化学習を行うので、高速に学習を行うことができ、さらに、環境に働きかける際には、その離散状態を再度連続状態へと変換しているので、連続時間及び連続状態を取り扱うことができるとともに、環境から自律的且つ高速に学習を行うことができる。

【発明を実施するための最良の形態】

【００１８】

まず、本発明の基本原理について説明する。階層型強化学習の目的の一つは、抽象化された状態空間で行動選択を行う上位層を設けることにより学習時間を短縮することである。このような上位層を用いる場合、「どのように状態や行動を抽象化するか」が最も難しく、また学習装置の性能に大きく関わってくる。特に、非線形な連続システムでは、環境の特性を実験者が把握することは困難なため、この抽象化が難しい。したがって、実験者を介さず、自律的（自動的）に抽象化が行える手法が必要となる。このようなモデルのひとつとして、WolpertらによるＭＯＳＡＩＣ（MOdule Switching And Identification for Control）モデルが挙げられる。

【００１９】

強化学習におけるＭＯＳＡＩＣモデルとは、環境に対して働きかけ、その結果得られる報酬を最大化するための行動出力を決定する強化学習システムと、環境の変化を予測する環境予測システムとの組によりなる学習モジュールが複数備えられ、各学習モジュールの環境予測システムの予測誤差が少ないものほど大きな値を取る責任信号が求められ、この責任信号に比例して強化学習システムによる制御出力（行動出力）が重み付けされて、環境に対する行動が与えられることを特徴とする学習モデルのことである（特許第３０８６２０６号参照）。

【００２０】

上記のＭＯＳＡＩＣモデルは、環境に対する制御出力と、環境を観測することで得られる観測変数との入出力特性に基づいて、環境の文節化を自動で行うことができる。そして、分けられた文節ごとに制御器を切り替えることにより、非線形又は非定常な環境への適応を可能としている。

【００２１】

次に、上記のＭＯＳＡＩＣモデルに基づいた環境の抽象化について説明する。本明細書で対象とする環境においては、ある時刻ｔにおける環境の状態ｘ（ｔ）（次元をＤｓとする）と、エージェント学習装置が環境に働きかける際の制御出力ｕ（ｔ）（次元をＤｃとする）と、制御出力ｕ（ｔ）のもとでの環境の状態変化ｘ’（ｔ）と、エージェント学習装置が環境を観測することにより得られる観測変数ｙ（ｔ）＝ｇ（ｘ（ｔ））＋ν（ｔ）（ν（ｔ）は観測ノイズ）との同時分布が、以下のように分解できると仮定する。なお、本明細書においては、文字に付けられたダッシュ（プライム）記号「’」は、文字の上に付けられた点（ドット）「・」を意味するものとする。ただし、文字としては、ｉ，ｊ，Ｓを除くものとし、ｉ，ｊに付けられたダッシュ記号はそのまま、そのものを意味し、Ｓ

’はｕによる偏微分を表すものとする。

【００２２】

【数１】

$$p(\dot{x}(t),y(t) \mid x(t),u(t)) = \sum_{i=1}^{N^f} \lambda_i^f(t)\, p(y(t) \mid x(t),i)\, p(\dot{x}(t) \mid x(t),u(t),i) \quad \cdots (1)$$

【００２３】

ここで、ｐ（ｘ’（ｔ）｜ｘ（ｔ），ｕ（ｔ），ｉ）は局所的なダイナミクスを表し、N^fはその個数、λ_i^f（ｔ）は前述の責任信号と呼ばれる重み付けのための信号の中のｉ番目の信号である。以下同様に、ｉ等の下付の添字は、その添字が付けられた変数や状態等のｉ番目のものを表すものとする。また、ｇは観測関数と呼ばれ、エージェント学習装置はこの観測関数ｇを通じて環境の状態ｘを獲得することができる。

【００２４】

一般に、複数個の局所ダイナミクスが切り替わる環境のモデルは、“Switching State-Space Models （ＳＳＭ）”と呼ばれる。本発明においては、ＭＯＳＡＩＣモデルに基づいて、局所ダイナミクスが切り替わるＳＳＭとして環境をモデル化し、各時刻において有効になっている局所ダイナミクスのインデックスを上位層の状態として採用する。

【００２５】

ＭＯＳＡＩＣモデルに基づいた環境の文節化は、観測変数ｙ（ｔ）の時系列から現在の状態ｘ（ｔ）を推定し、状態予測モデルの予測誤差に基づいて責任信号λ_i^f（ｔ）を求めることにより実現される。ここで、状態予測モデルとは、状態変化ｘ’（ｔ）の予測値を出力するモデルであり、ω（ｔ）をシステムノイズとして以下の式で表される。

【００２６】

【数２】

$$p(\hat{\dot{x}}(t) \mid x(t),u(t),i) = f_i(x(t),u(t)) + \omega(t) \quad \cdots (2)$$

【００２７】

図１（ａ）は後述するセミマルコフ決定過程（Semi-Markov Decision Processes、以下、ＳＭＤＰという）を模式的に示す図であり、図１（ｂ）は環境の状態変化を求めるための状態予測モデルを模式的に示す図である。図１（ｂ）は、連続状態空間において、環境の状態ｘ（ｔ）と制御出力ｕ（ｔ）とに基づいて、状態変化ｘ’（ｔ）を算出するダイナミクスを表している。この連続状態空間を複数の局所領域に分割し、それぞれにインデックスｉ＝１，２，．．．，N^fを割り当てると、図に示したように、そのインデックスｉにより指定される離散状態が形成される。図１（ａ）は、この離散状態間のダイナミクスの移り変わりを示しており、例えば、矢印の始点から終点へと離散状態が変化する様子を表している。

【００２８】

また、上記した状態ｘ（ｔ）の推定にはカルマンフィルタやモンテカルロ法などの手法を利用することができる。以下では、それらの手法により、状態ｘ（ｔ）の推定値ｘ~（ｔ）が得られたものとして責任信号の説明を行う。責任信号λ_i^f（ｔ）は、時刻ｔまでの状態（の推定値）ｘ~（ｔ）とその変化ｘ’~（ｔ）、及び制御出力ｕ（ｔ）の時系列の事後分布として以下の式に基づいて推定される。なお、本明細書においては、文字に付けられたティルダ「~」及びハット「^」は、文字の上に付けられていることを意味するものとする。そして、上記のｘのように「’」及び「~」、又は「’」及び「^」が２つ

重ねてある場合には、文字の上にドットが付けられ、さらにその上にティルダ又はハットが付けられているものとする。

【００２９】

【数３】

$$\lambda_i^f(t) \equiv P(i|\dot{\tilde{x}}[0:t], \tilde{x}[0:t], u[0:t]) \qquad \cdots (3)$$

【００３０】

責任信号λ$_i$^{f}（ｔ）は、各状態予測モデルｆ$_i$（ｘ（ｔ），ｕ（ｔ））の尤度を正規化することで求められるが、その尤度を状態予測誤差｜｜ｘ$_i$’＾（ｔ）－ｘ’～（ｔ）｜｜^{2}に基づいて計算するのが責任信号の特徴である。そのような文節化によって非線形・非定常な環境を適応的に制御できることがこれまでの研究で示されている（例えば、D. M. Wolpert and M. Kawato: “Multiple paired forward and inverse models for motor control”, Neural Networks, 11, pp. 1317-1369 (1998). 参照）。

【００３１】

したがって、各時刻において、最も高い責任信号を持つ状態予測モデルのインデックスｉは、環境を良く抽象化していると言え、そのインデックスｉを状態変数とする上位層はＭＯＳＡＩＣモデルの利点を有する状態空間で行動選択を行える。抽象化状態のインデックスをｉ、選択可能な行動（サブゴール）のインデックスをｊとすると、上位層が扱う環境は以下のようなＳＭＤＰとして定義できる。

【００３２】

１）状態ｉで行動ｊを選択したとき、次状態ｉ’は、ある状態遷移確率Ｐ（ｉ’｜ｉ，ｊ）に従って決定される。

【００３３】

２）状態ｉから次状態ｉ’に遷移するまでの時間ｌは、一定ではなく、ある確率分布Ｐ（ｌ｜ｉ’，ｉ，ｊ）を持つ。

【００３４】

３）次状態ｉ’に遷移したときに与えられる報酬は、状態ｉにいたときの報酬Ｒ（ｔ）を時間で積分した値となる。

【００３５】

このＳＭＤＰは、状態間の遷移や値の更新といった処理の時間間隔が一定ではなく、イベントドリブンすなわちそれら処理の時間間隔が任意な場合に対応した数理モデルである。また、上記したようなＳＭＤＰ環境において、どの抽象化状態が現時刻における環境の状態を最もよく表しているかを示す最適状態価値関数Ｖ*は、以下に示すＢｅｌｌｍａｎ方程式を満たさなければならない。

【００３６】

【数４】

$$V^*(i) = \max_j \left[\sum_{i'=1}^{N^f} P(i'|i,j) \int_0^{\infty} P(l|i',i,j) \left\{ \int_t^{t+l} e^{-\frac{s-t}{\tau}} R(s)ds + e^{-\frac{l}{\tau}} V^*(i') \right\} dl \right] \qquad \cdots (4)$$

【００３７】

ここで、右辺中括弧内の第一項は、時刻ｔからｔ＋ｌまで状態ｉに滞在したときに得る

報酬の重み付き累積和であり、τはその時定数である。また、右辺中括弧内の第二項中の指数部分は、次状態ｉ’に遷移するまでの時間ｌに応じた割引率である。そして、これら中括弧内の量を、状態遷移確率Ｐ（ｉ’｜ｉ，ｊ）と確率分布Ｐ（ｌ｜ｉ’，ｉ，ｊ）とで平均化している。このＳＭＤＰにおける定式化は、時間が連続で且つ遷移する時間が分布を持つことを除けばＭＤＰにおける定義と同様である。また、どの行動を取れば環境から得られる報酬を最も大きくすることができるかを示す最適行動価値関数Ｑ＊（ｉ，ｊ）は以下の式を満たす。

【００３８】

【数５】

$$Q^*(i,j)=\sum_{i'=1}^{N^f}P(i'|i,j)\int_0^{\infty}P(l|i',i,j)\left\{\int_t^{t+l}e^{-\frac{s-t}{\tau}}R(s)ds+e^{-\frac{l}{\tau}}\max_{j'}Q^*(i',j')\right\}dl \quad \cdots(5)$$

【００３９】

以上のように、ＭＯＳＡＩＣモデルに基づいたＳＳＭは、セミマルコフ的に局所ダイナミクスが切り替わることを特徴とする。

【００４０】

次に、本発明の一実施の形態によるエージェント学習装置について説明する。図２は、本発明の一実施の形態によるエージェント学習装置の構成を示すブロック図である。

【００４１】

図２に示すエージェント学習装置１は、上位層１０及び下位層２０の二層構造からなる。エージェント学習装置１は、ＲＯＭ（リードオンリメモリ）、ＣＰＵ（中央演算処理装置）、ＲＡＭ（ランダムアクセスメモリ）等を備える通常のマイクロコンピュータ、Ａ／Ｄ（アナログ／デジタル）変換器、Ｄ／Ａ（デジタル／アナログ）変換器等から構成され、ＲＯＭに記憶されたエージェント学習プログラムをＣＰＵにおいて実行することにより、上位層１０及び下位層２０として機能し、環境５０の状態を通じて学習を行う。

【００４２】

本実施の形態においては、ＭＯＳＡＩＣモデルが階層型強化学習装置に適用され、ＭＯＳＡＩＣモデルに基づいて切り分けられた文節を、行動選択を行う上位層に与える抽象化状態とすると、下位層２０は、環境状態の抽象化を行い、上位層１０は、各抽象化状態においてどの行動を目標とするべきかを強化学習により獲得する。そして、下位層２０は、上位層１０が選択した行動の達成度に応じて与えられる下位報酬を最大にするような制御則の学習を行う。

【００４３】

上位層１０は、行動選択器１０１、下位報酬関数選択器１０２、行動価値関数学習器１０３及び行動価値関数記憶器１０４を備え、行動選択を行う層であり、標準的な離散系強化学習の枠組みで記述される。そのため、ＭＯＳＡＩＣモデルも含め、これまでに離散系を対象として提案されている手法を容易に適用できる。

【００４４】

下位層２０は、内部状態推定器２０１、責任信号推定器２０２、制御信号出力器２０３、状態予測モデル記憶器２０４、下位状態価値関数学習器２０５及び下位状態価値関数記憶器２０６を備え、上位層１０と連続な状態をとる環境５０との間を取り持つ層である。なお、以下の説明では、行動選択器１０１が行う行動選択と行動価値関数学習器１０３が行う行動価値関数の更新とは、上位層１０に与えられる抽象化状態が変化したとき、つま

り最大責任信号を持つ状態予測モデルのインデックスｉが変化したときに行われるとする。また、上位層１０の状態が時刻ｔにてＩだったものが、時刻ｔ＋１にてＩ’へ変化したものとして説明する。

【００４５】

内部状態推定器２０１は、時刻０から時刻ｔまでの間に環境５０から受け取った観測変数ｙ［０：ｔ］と、時刻０から時刻ｔまでの間に制御信号出力器２０３から受け取った制御出力ｕ［０：ｔ］とから、現時刻ｔにおける環境５０の状態ｘ（ｔ）及びその変化（状態変化）ｘ’（ｔ）の推定値ｘ~（ｔ），ｘ’~（ｔ）を算出し、責任信号推定器２０２に送出する（例えば、R. E. Kalman: “A new approach to linear filtering and prediction problems”, Transaction of the ASME-Journal of Basic Engineering, 82, Series D, pp. 35-45 (1960). またはA. Doucet, N. de Freitas and N. Gordon: “Sequential Monte Carlo Methods in Practice”, Springer-Verlag (2001). 参照）。

【００４６】

ここで、ｖａ［０：ｔ］（ｖａは任意の変数を表す）は、時刻０から時刻ｔの間に観測された変数ｖａの閉区間における時系列である。また、ｖａ（０：ｔ）は、時刻０から時刻ｔの間に観測された変数ｖａの開区間における時系列である。

【００４７】

責任信号推定器２０２は、内部状態推定器２０１から受け取った環境５０の状態ｘ（ｔ）を抽象化する。つまり、責任信号推定器２０２は、複数の状態予測モデル｛f_i（ｘ（ｔ），ｕ（ｔ））｝を有する状態予測モデル記憶器２０４を参照し、制御の各時刻において責任信号λ_i^f（ｔ）を上位層１０に伝える。責任信号λ_i^f（ｔ）は、現時刻の情報から求められる各状態予測モデルf_i（ｘ（ｔ），ｕ（ｔ））の尤度と、それ以前の時系列から求められる事前予測値$\hat{\lambda}_i^f$（ｔ）との積に比例する。この責任信号は、時間が連続である場合、以下の式で定義される。

【００４８】

【数６】

$$\lambda_i^f(t) \equiv \frac{\hat{\lambda}_i^f(t)\, p(\dot{\tilde{x}}(t) \mid \tilde{x}(t), u(t), i)}{\sum_{i'=1}^{N^f} \hat{\lambda}_{i'}^f(t)\, p(\dot{\tilde{x}}(t) \mid \tilde{x}(t), u(t), i')} \quad \cdots (6)$$

$$\hat{\lambda}_i^f(t) \equiv \frac{P(i \mid \tilde{x}(t), u(t))\, P(i \mid \dot{\tilde{x}}[0:t), \tilde{x}[0:t), u[0:t))}{\sum_{i'=1}^{N^f} P(i' \mid \tilde{x}(t), u(t))\, P(i' \mid \dot{\tilde{x}}[0:t), \tilde{x}[0:t), u[0:t))} \quad \cdots (7)$$

【００４９】

ここで、確率分布ｐ（ｘ’~（ｔ）｜ｘ~（ｔ），ｕ（ｔ），ｉ）は状態予測モデルf_i（ｘ（ｔ），ｕ（ｔ））の尤度であり、予測誤差が分散$\sigma^{f\,2}$に従う正規分布だとした場合は、以下の式で定義される。

【００５０】

【数７】

$$p(\tilde{\dot{x}}(t)|\tilde{x}(t),u(t),i)=$$

$$\frac{1}{(2\pi)^{Ds/2}\sigma_f^{Ds}}\exp\left[-\frac{1}{2\sigma_f^2}\left\|\hat{\dot{x}}_i(t)-\tilde{\dot{x}}(t)\right\|^2\right]$$

・・・(8)

【００５１】

ここで、ｘ’＾（ｔ）はｉ番目の状態予測モデル$ｆ_{i}$（ｘ（ｔ），ｕ（ｔ））の予測出力値である。責任信号の事前予測値$λ^{\wedge}{}_{i}{}^{f}$（ｔ）は、空間的局所性と時間的連続性との２つの要素から得られ、式（７）においては、Ｐ（ｉ｜ｘ～（ｔ），ｕ（ｔ））が空間的局所性、Ｐ（ｉ｜ｘ’～［０：ｔ），ｘ～［０：ｔ），ｕ［０：ｔ））が時間的連続性に基づいた要素となっている。

【００５２】

以上のようにして推定される責任信号を用いて各状態予測モデルの出力や学習係数を重み付けすることにより、非線形且つ非定常なダイナミクスの各局所領域を各状態予測モデルが適応的に予測するような分化が行われる。特に、入力に責任信号で重み付けした値の分散を用いて空間的局所性を更新することで分化が促進される。

【００５３】

行動選択器１０１は、責任信号推定器２０２から責任信号$λ_{i}{}^{f}$（ｔ）の推定値を受け取り、当該責任信号を最大にするインデックスｉを決定（それをＩとする）し、当該インデックスＩを、現在時刻において最も環境をよく抽象化している一の状態を指定するインデックスとして選択する。上位層１０は、この抽象化状態Ｉの中で学習を行う。

【００５４】

また、行動選択器１０１は、行動価値関数記憶器１０４に記憶されている一つの行動価値関数Ｑ（ｉ，ｊ）を参照し、当該行動価値関数Ｑ（ｉ，ｊ）を責任信号$λ_{i}{}^{f}$（ｔ）で重み付けした値を用いて行動選択を行う。これは、観測にノイズや隠れ状態が存在するとき、責任信号$λ_{i}{}^{f}$（ｔ）は分布を持ち、信念状態の一種とみなされるからである。そして、行動選択器１０１は、新たに選択された（インデックスＩで指定される）抽象化状態が、前回選択された抽象化状態と同じか否か、つまり環境５０の抽象化状態が変化したか否かを判定する。その結果、抽象化状態が変化したと判定された場合には、行動選択器１０１は、行動価値関数学習器１０３に対して通知信号を送出する。

【００５５】

ある行動を選択する確率（行動選択確率）は、例えばＢｏｌｔｚｍａｎｎ（ボルツマン）分布に従うとすると、逆温度パラメータβを用いて以下の式で記述される。

【００５６】

【数８】

$$P(j|\lambda_i^f(t),i=1,\ldots,N^f)=\frac{\exp\left[\beta\ \Sigma_{i=1}^{N^f}\lambda_i^f(t)Q(i,j)\right]}{\Sigma_{j'}\exp\left[\beta\ \Sigma_{i=1}^{N^f}\lambda_i^f(t)Q(i,j')\right]}\quad\cdots(9)$$

【００５７】

行動選択器１０１は、式（９）の確率分布に従って確率的にｊの値を決定し、その値をエージェント学習装置１が取るべき一の行動を指定するインデックスＪ（＝ｊ）として選択する。そして、行動選択器１０１は、抽象化状態Ｉと、選択された行動を指定するインデックスＪとを下位報酬関数選択器１０２に送出する。

【００５８】

下位報酬関数選択器１０２は、抽象化状態を指定するインデックスｉと、行動を表すインデックスｊとによって指定される複数の下位報酬関数｛r_{ij}｝を有している。そして、下位報酬関数選択器１０２は、行動選択器１０１によって決定された抽象化状態を指定するインデックスＩと、行動選択器１０１によって選択された行動を表すインデックスＪとによって指定される下位報酬関数r_{IJ}を、下位状態価値関数学習器２０５に送出する。

【００５９】

上位層１０は、ある状態遷移確率Ｐ（ｉ’｜ｉ，ｊ）に従って遷移する抽象化状態を環境として学習を行う。その状態遷移確率は、各抽象化状態ｉに対する下位報酬関数r_{ij}（ｘ（ｔ），ｕ（ｔ））によって決定されるが、その状態遷移が確定的であるほど上位層１０は学習を行い易い。そこで、下位報酬関数r_{ij}（ｘ（ｔ），ｕ（ｔ））は、関数空間において、抽象化状態と行動とを表すインデックスにより指定される離散的な領域内に頂点を持つ凸関数を採用することが好ましい。このような下位報酬関数を下位層２０に与えれば、任意の抽象化状態への遷移確率をより確定的にすることができる。

【００６０】

状態予測モデル記憶器２０４は、状態予測モデルf_i（ｘ（ｔ），ｕ（ｔ））を記憶している。また、状態予測モデル記憶器２０４は、制御信号出力器２０３に対して、状態予測モデルf_iを制御出力ｕで偏微分した値を出力する。さらに、状態予測モデル記憶器２０４は、下位状態価値関数学習器２０５に対して、状態予測モデルf_iから算出された予測出力値ｘ’＾（ｔ）を出力する。

【００６１】

以下、状態予測モデルf_i（ｘ（ｔ），ｕ（ｔ））と、下位報酬関数r_{ij}（ｘ（ｔ），ｕ（ｔ））との組み合わせからなる関数を下位状態価値関数V_{ij}（ｘ（ｔ））という。この下位状態価値関数V_{ij}（ｘ（ｔ））の最適値である最適下位状態価値関数V^*_{ij}（ｘ（ｔ））は、ξを０＜ξ≦τを満たす定数として、以下のＢｅｌｌｍａｎ方程式で記述される。

【００６２】

【数９】

$$\frac{1}{\xi} V^*_{ij}(x(t)) = \max_u \left[r_{ij}(x(t),u(t)) + \frac{\partial V^*_{ij}(x(t))}{\partial x} f_i(x(t),u(t)) \right] \quad \cdots (10)$$

【００６３】

下位状態価値関数記憶器２０６は、下位状態価値関数Ｖ$_{ij}$（ｘ（ｔ））を記憶している。下位状態価値関数学習器２０５は、下位状態価値関数記憶器２０６に記憶されている下位状態価値関数Ｖ$_{ij}$を参照し、当該下位状態価値関数Ｖ$_{ij}$を前述の式（１０）で与えられる最適下位状態価値関数Ｖ＊$_{ij}$に近づけるための学習を行う。下位状態価値関数学習器２０５が行う学習は、以下の式に示す誤差を小さくするように行われる。

【００６４】

【数１０】

$$\delta_{ij}(t) = r_{ij}(x(t),u(t)) - \frac{1}{\tau} V_{ij}(x(t)) + \frac{\partial V_{ij}(x(t))}{\partial x} \dot{x}_i(t) \quad \cdots (11)$$

【００６５】

ここでδ$_{ij}$はＴＤ誤差と呼ばれ、当該学習はこのＴＤ誤差の自乗値が減少する方向へ式（１１）のパラメータを更新することで実現される（例えば、K. Doya: “Reinforcement learning in continuous time and space”, Neural Computation, 12, pp. 219-245 (2000). 参照）。そして、下位状態価値関数学習器２０５は、更新後の新たな下位状態価値関数Ｖ$_{ij}$を下位状態価値関数記憶器２０６に記憶させるとともに、Ｖ$_{ij}$を状態ｘで偏微分した値を制御信号出力器２０３に対して送出する。

【００６６】

また、最適下位状態価値関数Ｖ＊$_{ij}$（ｘ（ｔ））に従った最適制御出力ｕ＊$_{ij}$（ｘ（ｔ））は、学習が収束したときに得られるものであり、以下の式で表される。

【００６７】

【数１１】

$$u^*_{ij}(x(t)) = S'^{-1}\left(\frac{\partial f_i(x(t),u(t))^T}{\partial u} \frac{\partial V^*_{ij}(x(t))^T}{\partial x} \right) \quad \cdots (12)$$

【００６８】

ここで、Ｓ’（　）は出力のコスト関数を制御出力ｕで偏微分した関数であり、Ｓ’$^{-1}$（　）はその逆関数を表す。制御信号出力器２０３は、式（１２）によって算出された制御出力ｕ$_{IJ}$（ｘ（ｔ））を内部状態推定器２０１に送出するとともに、当該制御出力ｕ$_{IJ}$（ｘ（ｔ））を用いて環境５０に働きかける。つまり、下位層２０は、上位層１０

の状態と行動とに応じた下位価値関数を学習しながら制御出力を行う。

【００６９】

行動価値関数Ｑ（ｉ，ｊ）は、各抽象化状態に対する行動の評価値を表すものである。行動価値関数学習器１０３は、制御信号出力器２０３が与えた制御出力ｕ$_{IJ}$に応じて環境５０から受け取った報酬Ｒ（ｔ）を観測し、行動価値関数Ｑ（ｉ，ｊ）が持つ評価値の誤差を算出する。そして、行動価値関数学習器１０３は、抽象化状態が変化したことを知らせる通知信号を行動選択器１０１から受け取ると、行動価値関数記憶器１０４に記憶されている行動価値関数Ｑ（ｉ，ｊ）の誤差を修正する。行動価値関数Ｑ（ｉ，ｊ）は、報酬Ｒの将来に渡った重み付累積和によって行動を評価するものであり、エージェントはその評価が最大となるような行動則を学習する。行動価値関数Ｑ（ｉ，ｊ）は、式（５）で表される最適に近づくように学習が行われる。抽象化状態Ｉに対して選択された行動をＪ、変化した抽象化状態をＩ’とすると、学習は学習係数αを用いて以下の式で行われる。

【００７０】

【数１２】

$$Q(I,J):=Q(I,J)+\alpha\left[\int_t^{t+l} e^{-\frac{s-t}{\tau}} R(x(s),u(s))ds\right.$$

$$\left.+e^{-\frac{l}{\tau}}\max_j Q(I',j)-Q(I,J)\right] \quad \cdots(13)$$

【００７１】

ここで、τは０＜τ＜∞を満たす割引の時定数である。また、式（１３）における符号「：＝」は、当該符号の右辺の値を新たに左辺の値とする、更新処理を行うことを意味する。

【００７２】

本実施の形態においては、内部状態推定器２０１、責任信号推定器２０２及び行動選択器１０１が環境抽象化手段の一例に相当し、行動選択器１０１が状態決定手段及び行動決定手段の一例に相当し、下位報酬関数選択器１０２が下位報酬選択手段の一例に相当し、制御信号出力器２０３が制御出力決定手段の一例に相当する。

【００７３】

次に、上記のように構成されたエージェント学習装置の学習処理について説明する。図３は、図２に示すエージェント学習装置の学習処理を説明するためのフローチャートである。

【００７４】

まず、ステップＳ１において、内部状態推定器２０１は、環境５０から受け取った観測変数ｙの時系列と、制御信号出力器２０３から受け取った制御出力ｕとから、現時刻ｔにおける環境５０の状態ｘ（ｔ）及びその変化ｘ’（ｔ）の推定値ｘ~（ｔ），ｘ’~（ｔ）を算出する。内部状態推定器２０１から推定値ｘ~（ｔ），ｘ’~（ｔ）を受け取った責任信号推定器２０２は、ステップＳ２において、それらの値に基づき、時刻ｔにおける責任信号λ$_i^f$（ｔ）を推定する。

【００７５】

次に、ステップＳ３において、行動選択器１０１は、まず、責任信号推定器２０２から受け取った責任信号の推定値λ~$_i^f$（ｔ）を最大にするインデックスｉを決定し、当該インデックスｉを環境５０の抽象化状態Ｉとして選択する。そして、ステップＳ４において、行動選択器１０１は、新たに選択されたインデックスＩが、前回選択されたインデックスＩと同じか否か、つまり環境５０の抽象化状態が変化したか否かを判定する。その結

果、抽象化状態が変化していないと判定された場合（ステップＳ４でＮＯ）には、引き続きステップＳ７以降の処理が行われる。それに対して、抽象化状態が変化したと判定された場合（ステップＳ４でＹＥＳ）には、ステップＳ５以降の処理として新たな選択処理が行われる。

【００７６】

具体的には、ステップＳ５において、行動選択器１０１は、責任信号で重み付けされた行動価値関数Ｑを参照して、前述の式（９）の行動選択確率に従って確率的にｊの値を決定し、その値をエージェント学習装置１が取るべき行動Ｊとして選択する。そして、行動選択器１０１は、抽象化状態Ｉと、行動を指定するインデックスＪとを下位報酬関数選択器１０２に送出する。次に、下位報酬関数選択器１０２は、ステップＳ６において、抽象化状態Ｉと行動Ｊとによって指定される下位報酬関数r_{IJ}を下位状態価値関数学習器２０５に送出する。

【００７７】

次に、ステップＳ７において、下位報酬関数r_{IJ}を受け取った下位状態価値関数学習器２０５は、式（１１）で与えられるＴＤ誤差の自乗値を小さくするように下位状態価値関数を更新する。制御信号出力器２０３は、ステップＳ８において、下位状態価値関数$V_{ij}(x(t))$に従った制御出力u_{IJ}を決定し、内部状態推定器２０１へ送出するとともに、環境５０に対して出力することで環境５０に働きかける。

【００７８】

そして、行動価値関数学習器１０３は、与えられた制御出力u_{IJ}に応じた報酬Ｒを環境５０から受け取ると、ステップＳ９において、報酬Ｒの積分を行う。このとき、ステップＳ１０において、行動価値関数学習器１０３が行動選択器１０１から通知信号を受け取っていない場合、つまり抽象化状態が変化していないと判定された場合（ステップＳ１０でＮＯ）には、引き続きステップＳ１２以降の処理が行われる。それに対して、抽象化状態が変化したと判定された場合（ステップＳ１０でＹＥＳ）には、行動価値関数学習器１０３は当該行動価値関数による更新処理を行う。

【００７９】

つまり、上位層１０における学習の結果を反映させ、抽象化状態ｉが選択されたときにどの行動ｊを選択すれば最も価値があるか、つまり最も大きな報酬が得られるかを行動価値関数Ｑの更新という形で取り込む。具体的には、ステップＳ１１において、行動価値関数学習器１０３は、環境５０から与えられた報酬Ｒに基づいて算出された新たな行動価値関数Ｑにより、行動価値関数記憶器１０４に記憶された行動価値関数Ｑの値を更新する。そして、以上のステップＳ１からＳ１１までの処理を、学習が終了する（ステップＳ１２でＹＥＳ）まで繰り返し行う。

【００８０】

ここで、図３においては、ステップＳ４及びＳ１０で抽象化状態Ｉが変化した場合にのみ、ステップＳ５、Ｓ６及びＳ１１の処理を行うとして説明した。つまり、例えば、行動価値関数Ｑの更新は、一定の時間間隔ではなく、責任信号を最大にする状態が変わったときに（セミマルコフ的に）行われるとした。しかしながら、本実施の形態はそれに限られることなく、ステップＳ５、Ｓ６及びＳ１１の処理を、一定の時間間隔ごとに行う形態であってもよい。

【００８１】

上記の一連の処理により、本発明に係る実施の形態では、一般に連続状態を有する環境を離散状態へと抽象化し、当該離散状態において強化学習を行うので、高速に学習を行うことができる。また、環境に働きかける際には、その離散状態を再度連続状態へ変換しているので、連続時間及び連続状態を有する環境を取り扱うことが可能となり、当該環境から自律的且つ高速に学習を行うことができる。

【００８２】

次に、上記のエージェント学習装置の学習効果について、具体例を挙げて説明する。図４は、倒立振子の構成を示す図である。

【００８３】
図４に示すように、振子６０は、台車７０の一側面の中心に回動自在に取り付けられている。振子６０の位置は、当該振子６０が取り付けられた中心を通る鉛直線の上方の部分から測った角度θによって指定されるものとする。台車７０は、図４における右向きをｚ軸の正の向きとした場合に、ｚ軸に沿って加えられる力Ｆによって一軸方向に移動可能なように構成されている。
【００８４】
力Ｆの大きさ（Ｆの絶対値）｜Ｆ｜は、｜Ｆ｜＜２０［Ｎ］であるとする。また、台車７０の位置は、図の点線で示される位置を基準（ｚ＝０とする）とし、当該基準位置から振子６０が取り付けられた中心までのｚ座標で表すものとする。さらに、台車７０が移動可能な幅に制限はないものとする。
【００８５】
このような構成において、状態ｘ（ｔ）は、台車７０の座標ｚ、その座標ｚの変化ｚ’、角度θ、そしてその角度の変化θ’によって指定され、ｘ（ｔ）＝［ｚ　θ　ｚ’　θ’］^{T}（Ｔは行列の転置を表す）と表すことができる。また、制御出力ｕ（ｔ）は合力Ｆであるため、ｕ（ｔ）＝Ｆである。各観測関数は、g_i（ｘ（ｔ））＝ｘ（ｔ）＋Ｎ（０，０．１I_{Ds}）とした（ただし、I_{Ds}はＤｓ次元の単位行列を表し、Ｎはガウスノイズ（ホワイトノイズ）を表すものとする。）。物理パラメータは、Ｂａｒｔｏら（A. G. Barto, R. S. Sutton and C. W. Anderson: “Neuronlike adaptive elements that can solve difficult learning control problems”, IEEE Transactions on Systems Man and Cybernetics, 13, pp.834-846 (1983). 参照）が用いたものと同じ設定にし、制御の間隔は０．０１［ｓｅｃ］とした。また、Ｎ（０，０．１I_{Ds}）はシステムノイズであり、状態変数の次元Ｄｓは４、制御入力の次元Ｄｃは１である。
【００８６】
１試行の長さは１０［ｓｅｃ］とし、各試行の初期状態は振子６０がぶら下がった状態（θ＝π［ｒａｄ］）から±１５π／１８０［ｒａｄ］の範囲でランダムに選択した。報酬関数Ｒ（ｘ（ｔ），ｕ（ｔ））は、振子６０が倒立した状態（θ＝０［ｒａｄ］）の周りで大きく与えられる正の報酬と出力のコストの和で、次式に示すように与えた。
【００８７】
【数１３】

$$R(x(t),u(t))=\exp\left[-\frac{1}{2}\theta(t)^2\right]-0.0001F(t)^2 \quad \cdots (14)$$

【００８８】
上記の倒立振子を用い、図１に示すエージェント学習装置１による階層型強化学習アルゴリズム（以下、「ＳＳＲＬ」という）と、標準的な強化学習の２つの手法を用いて実験を行った。これら２つの手法の設定を、それぞれ以下に示す。
【００８９】
図１に示すエージェント学習装置１では、下位層の状態予測モデルf_iとして、以下のような線形モデルを計１００個用意した。
【００９０】

【数１４】

$$f_i(x,u)=A_i(x-x_i^f)+B_iu+c_i \quad \cdots(15)$$

$$P(i|\dot{x}[0:t),x[0:t),u[0:t))=\frac{1}{N^f} \quad \cdots(16)$$

$$P(i|x(t),u(t))=\frac{P(i,x(t),u(t))}{\sum_{i'=1}^{N^f}P(i',x(t),u(t))} \quad \cdots(17)$$

$$P(i,x(t),u(t))=\frac{1}{(2\pi)^{Ds/2}|\sum_i^f|^{Ds/2}}\exp\left[-(x(t)-x_i^f)^T\sum_i^{f-1}(x(t)-x_i^f)\right] \quad \cdots(18)$$

【００９１】
また、責任信号λ$_i^f$（ｔ）の事前予測値λ^$_i^f$（ｔ）は、空間的局所性に基づくもののみを用いた。予測誤差の分散σ$^{f\,2}$は０．１^{2}で固定とし、その他のパラメータＡ$_i$，Ｂ$_i$，ｃ$_i$，ｘ$_i^f$，Σ$_i^f$等は以下に示すＥＭアルゴリズムを用いて予め学習したものを実験に用いた。

【００９２】
まず、学習に用いたデータは［０　－π～π　－１０～１０　－２π～２π］^{T}の範囲で一様にランダムな初期状態ｘ（０）からガウスノイズを制御入力（ｕ（ｔ）＝［１０Ｎ（０，１）］）として与える事で生成した。そして、０．１［ｓｅｃ］の試行を計１０００回行った。

【００９３】
学習の対象としたパラメータとその初期値は以下のようにして設定した。式（１５）のような線形モデルを状態予測モデルとした場合、Ａ$_i$の上２行が［Ｏ$_{Ds/2}$　Ｉ$_{Ds/2}$］（ただし、Ｏ$_{Ds/2}$はＤｓ／２次元の正方なゼロ行列を表す。）、Ｂ$_i$の上２行とｃ$_i$の上半分の要素が０となることは明らかであるため、その他の要素のみ学習の対象とし、初期値は－０．１～０．１の間における一様な乱数として与えた。入力の分散の初期値は０．１Ｉ$_{Ds}$で与え、台車の横方向（ｚ方向）に関係する要素のみを固定し、出力の分散σ$^{f\,2}$は０．１^{2}で固定した。状態予測モデルの中心ｘ$_i^f$は、用意したデータの範囲で一様な乱数とした。

【００９４】
そして、以下の処理（Ｅステップ及びＭステップ）を、下記の式（１９）～（２１）を用いて各パラメータが十分に収束するまで繰り返した。
・Ｅステップ：各データに対するｉ番目の状態予測モデルの責任信号λ$_i^f$（ｔ）を求める。
・Ｍステップ：各パラメータを更新する。

【００９５】

【数１５】

$$x_i^f = \langle x \rangle_i / \langle 1 \rangle_i \quad \cdots (19)$$

$$\Sigma_i^f = \langle (x - x_i^f)(x - x_i^f)^T \rangle / \langle 1 \rangle_i \quad \cdots (20)$$

$$W_i = \langle \dot{x}[x^T u^T 1] \rangle \langle [x^T u^T 1]^T [x^T u^T 1] \rangle^{-1} \quad \cdots (21)$$

【００９６】

ここで、Ｅステップにおいて、W_iはパラメータA_i，B_i，c_iを横に並べたものである。つまり、W_i＝［A_i，B_i，c_i］である。また、＜＞$_i$は責任信号による重み付き平均を表し、例えば、＜ｈ（ｔ）＞$_i$＝（１／Ｔ）（$\int_{t=0}^{T}$ｈ（ｔ）λ_i^f（ｔ）ｄｔ）（$\int_{t=0}^{T}$はｔ＝０からｔ＝Ｔに亘る定積分を求めることを意味する積分記号であり、Ｔはデータの個数を表す。）である。

【００９７】

上位層が下位層に与える下位報酬関数は、獲得された状態予測モデルをもとにし、次のように自動的に設定した。ある抽象化状態ｉの中心x_i^fと、他の抽象化状態の中心との間の距離を調べ、その距離が短かった順に自分自身を含めた５つを選択する。そして、選ばれた各抽象化状態の中心x_i^fに頂点を持ち、その領域の逆共分散$(\Sigma_i^f)^{-1}$を傾きとした２次関数：$-(x(t)-x_i^f)^T(\Sigma_i^f)^{-1}(x(t)-x_i^f)$を下位報酬関数として設定した。

【００９８】

また、行動価値関数Ｑは、Ａｄａｐｔｉｖｅ　ＲＴＤＰ（Adaptive Real-Time Dynamic Programming；　例えば、A. G. Barto, S. J. Bradtke and S. P. Singh：“Learning to act using real-time dynamic programming”, Artificial Intelligence, 72, 1-2, pp.81-138 (1995). 参照 ）を用いて行い、行動選択はε－Ｇｒｅｅｄｙ（例えば、R. S. Sutton and A. G. Barto: “Reinforcement Learning ”, MIT Press (1998). 参照）を用い、１／１０の確率でランダムに行うとした。さらに、行動価値関数の時定数τは２とした。

【００９９】

下位価値関数V_{ij}は、すべての状態予測モデルf_iと、下位報酬関数r_{ij}との組み合わせに対して用意し、ＮＧＮｅｔ（例えば、J. Moody and C. J. Darken: “Fast learning in networks of locally-tuned processing units”, Neural Computation, 1, 2, pp. 281-294 (1989). 参照）を用いて学習した。下位価値関数V_{ij}は、１×５×５×５個の基底を状態予測モデルf_iの中心x_i^fの周りに配置し、exponential eligibility trace （例えば、K. Doya: “Reinforcement learning in continuous time and space”, Neural Computation, 12, pp. 219-245 (2000). 参照）によってパラメータを更新した。その際、学習係数αは０．１、下位価値関数の時定数ξは１．０、適格度の係数は０．５とした。下位層の出力ｕは、各時刻における上位層の状態と行動とに対応した状態予測モデルf_Iと下位価値関数V_{IJ}とを用いて次のように定義した。また、コスト関数は、Ｓ（Ｆ（ｔ））＝０．０００１Ｆ（ｔ）^{2}とした。

【０１００】

【数１６】

$$u(t)=S'^{-1}\left(\frac{\partial f_I(x(t),u(t))^T}{\partial u}\ \frac{\partial V_{IJ}(x(t))^T}{\partial x}\right)$$

・・・(22)

【０１０１】

一方、標準的な強化学習として、上記のＳＳＲＬとは異なり、階層構造を持たない標準的な強化学習手法で行動則の獲得を行った。価値関数の近似には、ＮＧＮｅｔを用いた。ＮＧＮｅｔの基底は、１×２０×２５×２５個を状態空間へ格子上に配置し、報酬Ｒ（ｔ）をもとに学習を行った。行動出力ｕ（ｔ）は、ＳＳＲＬの下位層と同じように、状態予測モデルと価値関数の勾配をもとに生成した。状態予測モデルは、ＳＳＲＬの場合と同じ条件で事前に学習したものを用いた。学習に用いた価値関数の学習係数、時定数とその適格度の係数及び基底の個数と分散は、ＳＳＲＬの下位層と同じものを用いて公平な条件になるように設定した。

【０１０２】

図５は、長さが１０００試行の実験を１０回行ったときの結果を示す。図５において、実線がＳＳＲＬ、破線が標準的な強化学習による結果を表し、図の横方向に実線又は破線で繋がれた各点は、それぞれＳＳＲＬ、標準的な強化学習による１０００試行毎の平均を表し、上下方向に描かれた実線及び破線の線分は、それぞれＳＳＲＬ、標準的な強化学習による標準偏差を表すものである。

【０１０３】

図５からわかるように、価値関数の学習に用いた関数近似器、状態予測モデルの性能は、本実施の形態のＳＳＲＬと標準的な強化学習とにおいて同じであるにもかかわらず、本実施の形態のＳＳＲＬが勝る結果となった。つまり、標準的な強化学習では、連続な状態空間全体に報酬が伝播するのを待たなければ振子６０が振り上げられないが、ＳＳＲＬでは、抽象化された状態空間で報酬を伝播させることができるため、標準的な強化学習に比べはるかに高速に学習を行えた。また、ＳＳＲＬでは早い段階で報酬が広く伝播していたが、標準的な強化学習の場合は初期状態の付近に報酬が伝播する前に実験が終了してしまい、ほとんど振子６０を振り上げることができなかった。その結果、最後の１００試行における振り上げの成功率は、ＳＳＲＬが３９．８７５［％］、標準的な強化学習が２．１２５［％］であった。

【０１０４】

公平な比較を行うために、ＳＳＲＬの実験においてＮＧＮｅｔを用いて下位価値関数V_{ij}の学習を行った。下位価値関数V_{ij}は、組になっている状態予測モデルf_iと下位報酬関数r_{ij}とを環境の代替とすることで、直接制御則u_{ij}（ｘ）を導くことができる。特に、今回の実験のように状態予測モデルf_iが線形であり、下位報酬関数r_{ij}が２次形式の場合はLinear Quadratic Controller （ＬＱＣ； 例えば、D. P. Bertsekas: “Dynamic programming and optimal control”, Athena Scientif (1995). 参照）として容易に求められる。本実施の形態の実用性を高めるためには、下位の行動則をＬＱＣとして与えた方が良く、今回の実験と同じ条件で下位の行動則をＬＱＣとした場合の成功率は、最初の１００試行の段階で９１．５［％］となった。

【０１０５】

以上説明したように、抽象化された空間での状態遷移は、ＳＭＤＰの枠組みで記述されるため、本実施の形態では、ＳＭＤＰ環境に対する既存の手法をそのまま上位層の学習則として適用可能である。この利点を生んだ大きな要因は、ＳＭＤＰ環境と連続システムである環境との受け渡しをしている下位層の存在である。つまり、下位層は、連続システム

をＳＭＤＰ環境に抽象化する役割を果たしている。

【０１０６】

また、連続変数で記述される事象と離散変数で記述される事象とが混在するシステムは、Hybrid Dynamical System と呼ばれ、非線形で大規模な制御対象をこのHybrid Dynamical System としてモデル化することにより、制御器の性能が向上することが知られている。本実施の形態においては、このHybrid Dynamical System を階層型強化学習により制御する枠組みを示したものであるとも言える。

【図面の簡単な説明】

【０１０７】

【図１】セミマルコフ決定過程を模式的に示す図および環境の状態変化を求めるための状態予測モデルを模式的に示す図である。

【図２】本発明の一実施の形態によるエージェント学習装置の構成を示すブロック図である。

【図３】図２に示すエージェント学習装置の学習処理を説明するためのフローチャートである。

【図４】倒立振子の構成を示す図である。

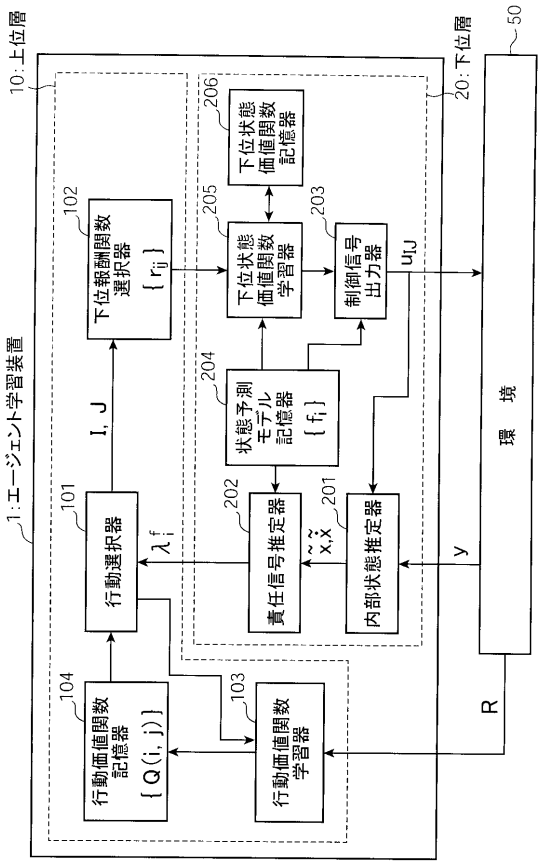
【図５】図４に示す倒立振子を用いた実験の結果を示す図である。

【符号の説明】

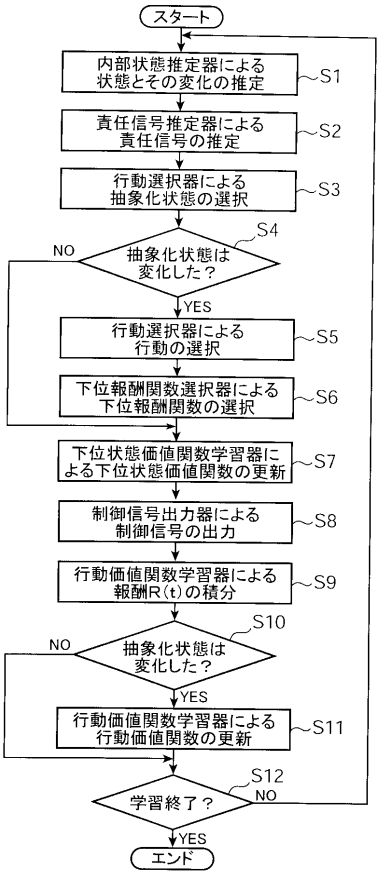
【０１０８】

１　エージェント学習装置
１０　上位層
２０　下位層
１０１　行動選択器
１０２　下位報酬関数選択器
１０３　行動価値関数学習器
１０４　行動価値関数記憶器
２０１　内部状態推定器
２０２　責任信号推定器
２０３　制御信号出力器
２０４　状態予測モデル記憶器
２０５　下位状態価値関数学習器
２０６　下位状態価値関数記憶器

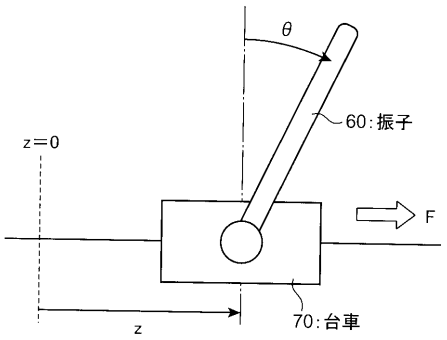
【図２】


10:上位層
20:下位層
50
1:エージェント学習装置
環境
102
下位報酬関数選択器
{ r_ij }
206
下位状態価値関数記憶器
205
下位状態価値関数学習器
203
制御信号出力器
u_iJ
204
状態予測モデル記憶器
{ f_i }
I, J
101
行動選択器
λ_i^f
202
責任信号推定器
201
内部状態推定器
y
R
104
行動価値関数記憶器
{ Q(i, j) }
103
行動価値関数学習器

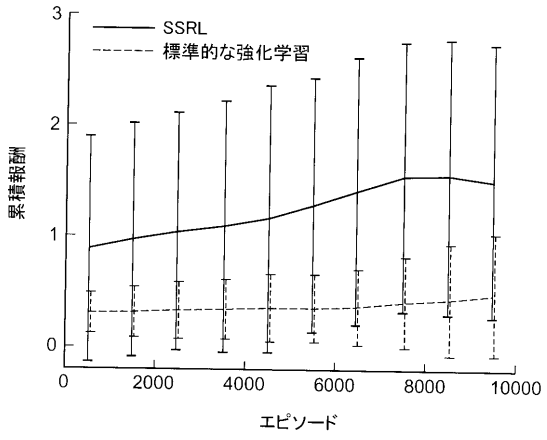
【図３】


スタート
内部状態推定器による状態とその変化の推定
S1
責任信号推定器による責任信号の推定
S2
行動選択器による抽象化状態の選択
S3
抽象化状態は変化した？
S4
NO
YES
行動選択器による行動の選択
S5
下位報酬関数選択器による下位報酬関数の選択
S6
下位状態価値関数学習器による下位状態価値関数の更新
S7
制御信号出力器による制御信号の出力
S8
行動価値関数学習器による報酬R(t)の積分
S9
抽象化状態は変化した？
S10
NO
YES
行動価値関数学習器による行動価値関数の更新
S11
学習終了？
S12
NO
YES
エンド

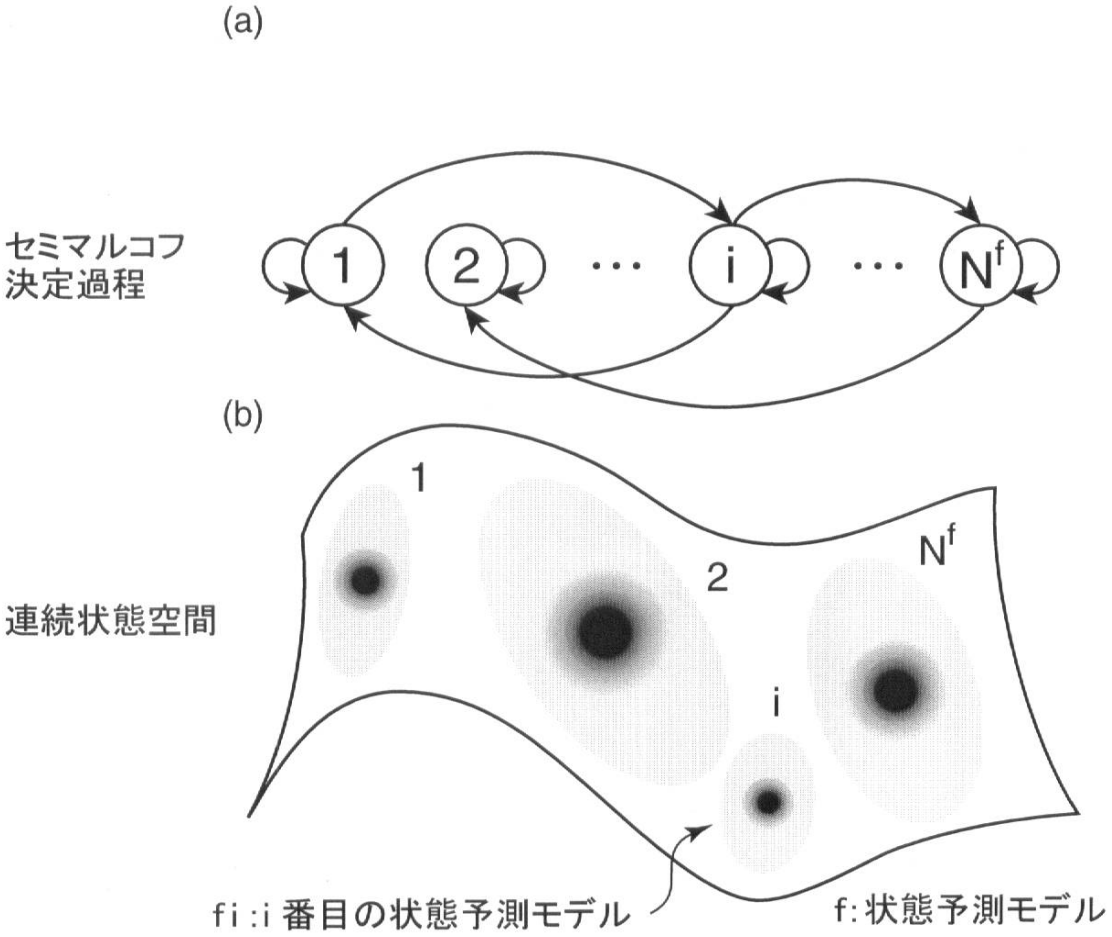
【図４】


θ
60:振子
z=0
F
z
70:台車

【図５】


SSRL
標準的な強化学習
累積報酬
エピソード
0
1
2
3
0
2000
4000
6000
8000
10000

【図１】



フロントページの続き

(72)発明者　銅谷　賢治
京都府相楽郡精華町光台二丁目２番地２　株式会社国際電気通信基礎技術研究所内
(72)発明者　川人　光男
京都府相楽郡精華町光台二丁目２番地２　株式会社国際電気通信基礎技術研究所内

審査官　長谷川　篤男

(56)参考文献　森本　淳、銅谷　賢治，階層型強化学習を用いた３リンク２関節ロボットによる起立運動の獲得，日本ロボット学会誌，２００１年　７月１５日，第19巻　第5号，第32-37頁
鮫島　和行、片桐　憲一、銅谷　賢治、川人　光男，複数の予測モデルを用いた強化学習による非線形制御，電子情報通信学会論文誌，２００１年　９月２０日，第J84-D-II巻 第9号，第2092-2106頁
高橋　泰岳、浅田　稔，階層型学習機構における状態行動空間の構成，日本ロボット学会誌，２００３年　３月１５日，第21巻　第2号，第38-45頁
鮫島　和行、銅谷　賢治、川人　光男，強化学習ＭＯＳＡＩＣ：予測性によるシンボル化と見まね学習，日本ロボット学会誌，２００１年　７月１５日，第19巻　第5号，第9-14頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　３／００
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）