(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2022-124284
(P2022-124284A)
(43)公開日　令和4年8月25日(2022.8.25)

(51)Int.Cl.		FI		テーマコード（参考）
G06N 20/00	(2019.01)	G06N 20/00		5L049
G06N 99/00	(2019.01)	G06N 99/00	180	
G06Q 10/04	(2012.01)	G06Q 10/04		

審査請求 未請求　請求項の数 4　OL　（全 12 頁）

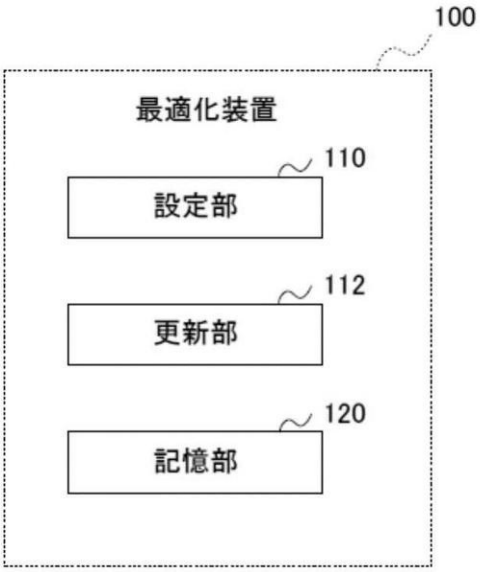
(21)出願番号　特願2021-21962(P2021-21962)
(22)出願日　令和3年2月15日(2021.2.15)

(71)出願人　501241645
学校法人　工学院大学
東京都新宿区西新宿１丁目２４番２号
(71)出願人　512155478
学校法人沖縄科学技術大学院大学学園
沖縄県国頭郡恩納村字谷茶１９１９番地１
(74)代理人　110001519
特許業務法人太陽国際特許事務所
(72)発明者　竹川　高志
東京都新宿区西新宿一丁目２４番２号　学校法人工学院大学内
(72)発明者　高橋　春輝
東京都新宿区西新宿一丁目２４番２号　学校法人工学院大学内

最終頁に続く

(54)【発明の名称】最適化装置、最適化方法、及び最適化プログラム

(57)【要約】
【課題】演算に係る効率を向上させることを可能とする。
【解決手段】隠れ状態を所定の態様に変更した独自隠れ状態、及び独自隠れ状態における現在の状態の推定を保持し、観測状態の法則は、条件付き確率の条件として時刻ｔの観測を用い、独自隠れ状態を得るように、状態の推移法則は、条件付き確率の条件として時刻ｔの独自隠れ状態及び時刻ｔ＋１の独自隠れ状態を用い、エージェントの行動を得るように、状態の推移法則、及び観測状態の法則を定義する。各法則を用いて、エージェントの手順に従って分布を更新する。
【選択図】図２



【特許請求の範囲】
【請求項１】
状態の推移法則、観測状態の法則、及び報酬の法則による各法則が定義されている系を用い、エージェントの行動を繰り返して前記各法則を学習し報酬を獲得するモデルにおいて、
隠れ状態を所定の態様に変更した独自隠れ状態、及び前記独自隠れ状態における現在の状態の推定を保持し、
前記観測状態の法則は、条件付き確率の条件として時刻ｔの観測を用い、前記独自隠れ状態を得るように、
前記状態の推移法則は、条件付き確率の条件として時刻ｔの前記独自隠れ状態及び時刻ｔ＋１の前記独自隠れ状態を用い、前記エージェントの行動を得るように、
前記状態の推移法則、及び前記観測状態の法則を定義する設定部と、
前記各法則をもとにサンプリングした確率を表す各パラメータの分布と、前記現在の状態の推定とを仮定して、ベルマン方程式に基づいて前記エージェントの最適行動を決定し、
前記各法則、所定の事前分布、及び前記最適行動を含む観測情報に対してベイズの定理を適用して得られた事後分布により、前記現在の状態の推定、及び前記各法則を用いた前記分布を更新することを繰り返す更新部と、
を含む最適化装置。
【請求項２】
前記更新部は、エージェントの動作の手順において、各パラメータをサンプリングすることにより、ベルマン方程式に基づき、長期報酬ｑが最大の行動を選択する請求項１に記載の最適化装置。
【請求項３】
状態の推移法則、観測状態の法則、及び報酬の法則による各法則が定義されている系を用い、エージェントの行動を繰り返して前記各法則を学習し報酬を獲得するモデルにおいて、
隠れ状態を所定の態様に変更した独自隠れ状態、及び前記独自隠れ状態における現在の状態の推定を保持し、
前記観測状態の法則は、条件付き確率の条件として時刻ｔの観測を用い、前記独自隠れ状態を得るように、
前記状態の推移法則は、条件付き確率の条件として時刻ｔの前記独自隠れ状態及び時刻ｔ＋１の前記独自隠れ状態を用い、前記エージェントの行動を得るように、
前記状態の推移法則、及び前記観測状態の法則を定義し、
前記各法則をもとにサンプリングした確率を表す各パラメータの分布と、前記現在の状態の推定とを仮定して、ベルマン方程式に基づいて前記エージェントの最適行動を決定し、
前記各法則、所定の事前分布、及び前記最適行動を含む観測情報に対してベイズの定理を適用して得られた事後分布により、前記現在の状態の推定、及び前記各法則を用いた前記分布を更新することを繰り返す、
処理をコンピュータに実行させる最適化方法。
【請求項４】
コンピュータを、請求項１又は請求項２に記載の最適化装置の各部として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、最適化装置、最適化方法、及び最適化プログラムに関する。
【背景技術】
【０００２】

問題設定に対する解決手法のアプローチとして、隠れマルコフモデル、ベイズ推定、及び強化学習等の手法が用いられている。

【０００３】

例えば、特許文献１には、隠れ状態数および観測確率の種類と共にモデルの候補数が指数的に増加しても高速にモデル選択を実現できる隠れ変数モデル推定装置が開示されている。この隠れ変数モデル推定装置は、周辺化対数尤度関数を完全変数に対する推定量に関してラプラス近似した近似量の下界として定義される基準値を最大化することによって変分確率を計算する変分確率計算部を有する。また、隠れ変数モデル推定装置は、各隠れ状態に対して観測確率の種類とパラメータを推定することで最適な隠れ変数モデルを推定するモデル推定部と、変分確率計算部が変分確率を計算する際に用いた基準値が収束したか否かを判定する収束判定部とを有する。

【０００４】

また、特許文献２には、環境と相互作用する強化学習エージェントが遂行する行動を選択するシステムが開示されている。このシステムは、目標回帰型ニューラルネットワーク（ＮＮ）の現在の隠れ状態に従って処理して、時間ステップについて、目標空間における初期の目標ベクトルを生成し、目標回帰型ＮＮの内部状態を更新するように構成される、処理する工程を有している。

【０００５】

また、強化学習は、状態と行動の組み合わせに対して報酬と次の状態が決定する手法である。

【先行技術文献】

【特許文献】

【０００６】

【特許文献１】再表２０１３／１７９５７９号公報

【特許文献２】特表２０２０－５０８５２４号公報

【発明の概要】

【発明が解決しようとする課題】

【０００７】

強化学習の枠組みにおいて、標準的にＱ学習と呼ばれる手法が用いられている。Ｑ学習は離散の状態に対して定義されるが、現実の課題は膨大な観測状態が存在するため、通常のＱ学習では学習が難しい場合が多い。

【０００８】

近年、発展系であるＱ学習と多層のニューラルネットワークを組み合わせたＤｅｅｐ　Ｑ　Ｎｅｔｗｏｒｋ　（ＤＱＮ）がさまざまな課題において有効であることが示されている。学習済みのＤＱＮは非常に高い性能を示すが、動作の内部状態がブラックボックスで与えられた環境をどのように解釈しているかが不明である。また、学習後の性能は高いが学習には多くの反復を必要とし，学習中に効果的に報酬を獲得することはあまり考慮されていない。

【０００９】

一方、膨大な観測から重要な隠れ状態を推定しつつ状態遷移を効果的に学習するベイズ推定を用いたアルゴリズムも広く知られている。しかし、この手法では報酬の予測と状態遷移を別に扱うため、報酬と無関係な状態を詳細に分析していることとなり、問題設定によってはメモリ、及び計算量などに多大な無駄が生じる。また、多腕バンディット問題と呼ばれる枠組みにおいて、学習と報酬獲得とをバランス良く行う汎用の手法としてトンプソンサンプリングが知られているが、複雑な問題に直接適用することはできない。

【００１０】

本発明は、上記事情を鑑みて成されたものであり、演算に係る効率を向上させることを可能とする最適化装置、最適化方法、及び最適化プログラムを提供することを目的とする。

【課題を解決するための手段】

【００１１】
上記目的を達成するために、本発明に係る最適化装置は、状態の推移法則、観測状態の法則、及び報酬の法則による各法則が定義されている系を用い、エージェントの行動を繰り返して前記各法則を学習し報酬を獲得するモデルにおいて、隠れ状態を所定の態様に変更した独自隠れ状態、及び前記独自隠れ状態における現在の状態の推定を保持し、前記観測状態の法則は、条件付き確率の条件として時刻ｔの観測を用い、前記独自隠れ状態を得るように、前記状態の推移法則は、条件付き確率の条件として時刻ｔの前記独自隠れ状態及び時刻ｔ＋１の前記独自隠れ状態を用い、前記エージェントの行動を得るように、前記状態の推移法則、及び前記観測状態の法則を定義する設定部と、前記各法則をもとにサンプリングした確率を表す各パラメータの分布と、前記現在の状態の推定とを仮定して、ベルマン方程式に基づいて前記エージェントの最適行動を決定し、前記各法則、所定の事前分布、及び前記最適行動を含む観測情報に対してベイズの定理を適用して得られた事後分布により、前記現在の状態の推定、及び前記各法則を用いた前記分布を更新することを繰り返す更新部と、を含んで構成されている。
【００１２】
本発明に係る最適化方法は、状態の推移法則、観測状態の法則、及び報酬の法則による各法則が定義されている系を用い、エージェントの行動を繰り返して前記各法則を学習し報酬を獲得するモデルにおいて、隠れ状態を所定の態様に変更した独自隠れ状態、及び前記独自隠れ状態における現在の状態の推定を保持し、前記観測状態の法則は、条件付き確率の条件として時刻ｔの観測を用い、前記独自隠れ状態を得るように、前記状態の推移法則は、条件付き確率の条件として時刻ｔの前記独自隠れ状態及び時刻ｔ＋１の前記独自隠れ状態を用い、前記エージェントの行動を得るように、前記状態の推移法則、及び前記観測状態の法則を定義し、前記各法則をもとにサンプリングした確率を表す各パラメータの分布と、前記現在の状態の推定とを仮定して、ベルマン方程式に基づいて前記エージェントの最適行動を決定し、前記各法則、所定の事前分布、及び前記最適行動を含む観測情報に対してベイズの定理を適用して得られた事後分布により、前記現在の状態の推定、及び前記各法則を用いた前記分布を更新することを繰り返す、処理をコンピュータに実行させる。
【発明の効果】
【００１３】
本発明の最適化装置、最適化方法、及び最適化プログラムによれば、演算に係る効率を向上させることを可能とする、という効果が得られる。
【図面の簡単な説明】
【００１４】
【図１】状態及び法則の推定に関して、従来手法の遷移図と、本実施形態の手法の遷移図との一例を示した図である。
【図２】本発明の実施形態に係る最適化装置の各機能構成を示す図である。
【図３】最適化装置のハードウェア構成を示すブロック図である。
【図４】本発明の実施形態に係る最適化装置の最適化処理ルーチンを示す図である。
【図５】本実施形態の手法と他の手法の実験結果の一例を示すグラフである。
【図６】実験における収束時の隠れ状態数を表にした図である。
【発明を実施するための形態】
【００１５】
以下、図面を参照して本発明の実施形態を詳細に説明する。
【００１６】
まず、本発明の実施形態における原理的な説明をする。
【００１７】
図１は、状態及び法則の推定に関して、従来手法の遷移図と、本実施形態の手法の遷移図との一例を示した図である。まず基本的な原理として、従来手法の状態及び法則の推定について説明する。従来手法、及び本実施形態の手法は共通して、状態の推移法則、観測

状態の法則、及び報酬の法則による各法則が定義されている系（遷移図）を用い、エージェントの行動を繰り返して各法則を学習し報酬を獲得する内部モデルを持つ。図１上は、従来手法の状態及び法則の推定の遷移図である。時刻ｔに対して観測o_tが得られ、行動a_tを選択すると報酬r_tと次の観測o_{t+1}が得られる環境が与えられたとする。この場合に、割引率γに対する長期報酬$\Sigma_{\tau=0}^{\infty}\gamma^{\tau}r_{t+\tau}$をできるだけ大きくするような選択を行いたい。また、観測o_tの背景には隠れ状態が存在し、行動によって隠れ状態が確率的に変化し、報酬も確率的に決定されるものとする。標準的には、隠れ状態をs_tとし，状態の推移法則$p(s_{t+1}|s_t, a_t)$、観測状態の法則$p(o_t|s_t)$及び報酬の法則$p(r_t|s_t, a_t)$が定義されている系を想定する。以下、推移法則、観測法則、及び報酬法則という。ただし、目的を達成するためのアルゴリズム（エージェント）にとってこれらの法則は未知であり。行動を繰り返して法則を学習しつつ並行して高い報酬を獲得する必要がある。

【００１８】

一方、本発明の実施形態に係る原理において、エージェントは、環境に対する内部モデルとして、独自隠れ状態s_t'と現在の状態の推定$q(s_t')$を保持している。ここで、s_t'はs_tの報酬に関連する要素に着目して簡略化したものを想定している。ただし、s_tは存在そのものと推移則は仮定しているものの実際に推定するわけではない。何らかの複雑な状態と推移則があるとして、それを直接考えることなく、報酬の観点から不要な状態を排除したものがs_t'である。

【００１９】

本実施形態の手法では、推移法則、観測法則、及び報酬法則についても内部モデルを持つが、観測法則と推移法則とに関して実際の法則と異なる$p(s_t'|o_t)$と$p(a_t|s_t', s_{t+1}')$との形式を用いていることが特徴である。報酬法則$p(r_t|s_t', a_t)$に関しては実際の法則と同様である。このような形式により、本実施形態の状態及び法則の推定の遷移図は、図１下のようにできる。エージェントは各法則のパラメータを確率分布として保持し、観測結果に応じて学習する。具体的には、確率を表すパラメータである$M'(s_t', s_{t+1}', a_t)=p(a_t|s_t', s_{t+1}')$，$N'(s_t', o_t)=p(o_t|s_t')$，$L(s_t', a_t, r_t)=p(r_t|s_t', a_t)$に対して、パラメータの予測である$q(M')$，$q(N')$，$q(L)$が設定されている。$q(M')$と$q(N')$は容易に実際の法則と対応する$q(M)$と$q(N)$に変換できる。

【００２０】

ここで、パラメータについて説明する。例えば、$p(r|s, L)$では、ｓ１という状態でｒの取り得る値がｒ１，ｒ２，ｒ３だったとすると、$L(s1, r1)$はｒ１が得られる確率を表し、$L(s1, r1)+L(s1, r2)+L(s1, r3)=1$と、Ｌは行列として表現できる。Ｍについては、ｓ，ｓ'，ａのｉｎｄｅｘをとるテンソルとなり、ａについて和をとると１となる。

【００２１】

エージェントは、次の手順で動作する。［１］確率分布$q(M)$，$q(N)$，$q(L)$に従ってＭ，Ｎ，Ｌをサンプリングする。［２］サンプリングしたＭ，Ｎ，Ｌと状態推定$q(s_t')$が正しいと仮定した場合の最適行動a_tをベルマン方程式に基づいて決定し出力する。最適行動a_tを出力した結果、新しい情報としてr_t，o_{t+1}を得る。［３］法則$p(a_t|s_t', s_{t+1}')$，$p(o_t|s_t')$，$p(r_t|s_t', a_t)$と、事前分布$q(s_t')$，$q(M')$，$q(N)$，$q(L)$と、観測された情報o_t，a_t，r_t，o_{t+1}に対してベイズの定理を適用する。得られた事後分布を、新たな知識$q(s_{t+1}')$，$q(M')$，$q(N)$，$q(L)$として更新する。［４］その後、［１］に戻り反復する。

【００２２】

ベルマン方程式では、行動ａを行った場合の長期報酬ｑの期待値が求まるので、単純にｑが最大の行動を取る。ベイズの定理を適用とは、各法則$p(a|s, s', M)$，$p($

ｓ｜ｏ，Ｌ），ｐ（ｒ｜ｓ，Ｎ）と事前分布ｐ（Ｌ，Ｍ，Ｎ）を用いて、事後分布ｐ（ｓ，ｓ'，Ｌ，Ｍ，Ｎ｜ｏ，ｏ'，ｒ，ｒ'）を計算することを指す。以下に事後分布の計算の適用例を示す。

【００２３】

$$p(s, s', L, M, N \mid o, o', r, r', a)$$
$$\propto p(r, r', a, s, s', L, M, N \mid o, o')$$
$$= p(r \mid s, N)\, p(r' \mid s', N)\, p(a \mid s, s', M)\, p(s \mid o, L)\, p(s' \mid o', L)\, p(L, M, N)$$

【００２４】

その他、計算の手法は毎回変分ベイズを用いて事後分布を収束するまで計算するが、従来手法では毎回事後分布を収束するまで計算せず変分ベイズのｓｔｅｐのみ更新する実装がされている。また、本実施形態の手法の方が計算量は増えるが、オンライン性の大きな向上が見込める。

【００２５】

上記手順の特徴について説明する。手順［１］は状態推移についてトンプソンサンプリングを適用することが従来手法では試みられていない。従来手法としては、［１］及び［２］をまとめる形でニューラルネットワークによるｑの予測を行うことが主流である。また、手順［２］については、Ｍ，Ｎ，Ｌが既知の場合にｑを求めることは標準的な手法であり、Ｑを元にｓｏｆｔｍａｘで確率的に行動を決定するのが標準的な手法である。これに対して、本実施形態の手法では［１］でサンプリングしているので、ｓｏｆｔｍａｘは使わず単純にｑが最大の行動を選択する点に特徴がある。手順［３］については、従来手法では確率モデル$p(s_{t+1} \mid s_t, a_t)$，$p(o_t \mid s_t)$を仮定するのに対し、本実施形態の手法では、$p(a_t \mid s_t', s_{t+1}')$，$p(s_t' \mid o_t)$を仮定して定式化している点が大きく異なる。また、従来手法では観測o_tを単純なカテゴリカル分布と仮定しているのに対し、本実施形態の手法ではカテゴリカル分布の直積に拡張している。

【００２６】

なお、上記の例では、計算には変分ベイズを用いることとしているが、手順自体に変分ベイズが必須ではなく、他の計算手法を用いてもよい。

【００２７】

＜本発明の実施形態に係る最適化装置の構成＞

次に、本発明の実施形態に係る最適化装置の構成について説明する。

【００２８】

図２は、本発明の実施形態に係る最適化装置１００の各機能構成を示す図である。図２に示すように、この最適化装置１００は、機能的には、設定部１１０と、更新部１１２と、記憶部１２０とを備えている。

【００２９】

設定部１１０は、状態の推移法則、及び観測状態の法則の定義を設定し、当該設定を記憶部１２０に保存する。以下、各法則に関して、適宜当該設定を読み出して処理を行う。

【００３０】

上記原理において示したように、設定によって、隠れ状態s_tを所定の態様に変更した独自隠れ状態s_t'、及び独自隠れ状態における現在の状態の推定$q(s_t')$を保持する。設定によって、観測状態の法則は、条件付き確率の条件として時刻ｔの観測o_tを用い、独自隠れ状態s_t'を得るようにする（$p(s_t' \mid o_t)$）。設定によって、状態の推移法則は、条件付き確率の条件として時刻ｔの独自隠れ状態s_t'及び時刻ｔ＋１の独自隠れ状態s_{t+1}'を用い、エージェントの行動a_tを得るようにする（$p(a_t \mid s_t', s_{t+1}')$）。

【００３１】

更新部１１２は、エージェントの最適行動a_tを決定し、分布を更新することを繰り返す。更新は、予め定めた条件を満たすまで繰り返せばよい。エージェントの最適行動a_t

は、まず、上記エージェントの動作の手順［１］に従って、各法則をもとに確率を表す各パラメータのＭ，Ｎ，Ｌをサンプリングする。次に手順［２］に従って、サンプリングした各パラメータＭ，Ｎ，Ｌと、現在の状態の推定ｑ（s_t'）とを仮定して、ベルマン方程式に基づいてエージェントの最適行動a_tを決定し出力する。最適行動a_tを出力した結果、新しい情報としてr_t，o_{t+1}を得る。手順［３］に従って、各法則ｐ（a_t｜s_t'，s_{t+1}'），ｐ（o_t｜s_t'），ｐ（r_t｜s_t'，a_t）、所定の事前分布ｑ（s_t'），ｑ（Ｍ'），ｑ（Ｎ），ｑ（Ｌ）、及び最適行動を含む観測情報o_t，a_t，r_t，o_{t+1}に対してベイズの定理を適用して事後分布を得る。そして、更新部１１２は、得られた事後分布により、現在の状態の推定、及び各法則を用いた分布ｑ（s_{t+1}'），ｑ（Ｍ'），ｑ（Ｎ），ｑ（Ｌ）を更新することを繰り返すことにより、最終的に収束された分布を出力する。

【００３２】

記憶部１２０には、設定部１１０で設定された各法則に係る設定、更新部１１２の計算過程の計算データ、及び計算結果が保存される。

【００３３】

図３は、最適化装置１００のハードウェア構成を示すブロック図である。

【００３４】

図３に示すように、最適化装置１００は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）１１、ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）１２、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）１３、ストレージ１４、入力部１５、表示部１６及び通信インタフェース（Ｉ／Ｆ）１７を有する。各構成は、バス１９を介して相互に通信可能に接続されている。

【００３５】

ＣＰＵ１１は、中央演算処理ユニットであり、各種プログラムを実行したり、各部を制御したりする。すなわち、ＣＰＵ１１は、ＲＯＭ１２又はストレージ１４からプログラムを読み出し、ＲＡＭ１３を作業領域としてプログラムを実行する。ＣＰＵ１１は、ＲＯＭ１２又はストレージ１４に記憶されているプログラムに従って、上記各構成の制御及び各種の演算処理を行う。本実施形態では、ＲＯＭ１２又はストレージ１４には、最適化プログラムが格納されている。

【００３６】

ＲＯＭ１２は、各種プログラム及び各種データを格納する。ＲＡＭ１３は、作業領域として一時的にプログラム又はデータを記憶する。ストレージ１４は、ＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）又はＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）等の記憶装置により構成され、オペレーティングシステムを含む各種プログラム、及び各種データを格納する。

【００３７】

入力部１５は、マウス等のポインティングデバイス、及びキーボードを含み、各種の入力を行うために使用される。

【００３８】

表示部１６は、例えば、液晶ディスプレイであり、各種の情報を表示する。表示部１６は、タッチパネル方式を採用して、入力部１５として機能してもよい。

【００３９】

通信インタフェース１７は、端末等の他の機器と通信するためのインタフェースであり、例えば、イーサネット（登録商標）、ＦＤＤＩ、Ｗｉ－Ｆｉ（登録商標）等の規格が用いられる。

【００４０】

＜本発明の実施形態に係る最適化装置の作用＞

次に、本発明の実施形態に係る最適化装置１００の作用について説明する。最適化装置１００の各部としてＣＰＵ１１が、図４に示す最適化処理ルーチンを実行する。

【００４１】

ステップＳ１００では、ＣＰＵ１１が、状態の推移法則、及び観測状態の法則の定義を設定し、当該設定を記憶部１２０に保存する。

【００４２】

ステップＳ１０２では、ＣＰＵ１１が、エージェントの動作の手順［１］に従って、各パラメータのＭ，Ｎ，Ｌをサンプリングする。

【００４３】

ステップＳ１０４では、ＣＰＵ１１が、手順［２］に従って、サンプリングした各パラメータＭ，Ｎ，Ｌと、現在の状態の推定ｑ（s_t'）とを仮定して、ベルマン方程式に基づいてエージェントの最適行動a_tを決定する。

【００４４】

ステップＳ１０５では、ＣＰＵ１１が、最適行動a_tを出力した結果、新しい情報としてr_t，o_{t+1}を得る。

【００４５】

ステップＳ１０６では、ＣＰＵ１１が、手順［３］に従って、各法則、所定の事前分布、及び最適行動a_tを含む観測情報に対してベイズの定理を適用して事後分布を得る。

【００４６】

ステップＳ１０８では、ＣＰＵ１１が、更新の条件を満たすか否かを判定する。条件を満たすと判定した場合にはステップＳ１１０へ移行し、条件を満たさないと判定した場合にはステップＳ１０２に戻って処理を繰り返す（「手順［４］）。

【００４７】

ステップＳ１１０では、最終的に得られた分布を出力し、処理を終了する。

【００４８】

以上、説明した本発明の実施形態によれば、演算に係る効率を向上させることが可能である。また、技術のポイントは大きく、３つのポイントが挙げられる。１点目は確率モデルによる状態推定と意思決定問題とを統合した点、２点目は観測則と推移則とを通常の形式でなく独自の形式の法則とした点、３点目は手順におけるサンプリングの活用である。

【００４９】

１点目の確率モデルによる状態推定と意思決定問題とを統合した点について説明する。これまで、カルマンフィルタなど観測からの隠れ状態の推定モデルについては様々な手法が提案されている。また、Ｑ学習を代表として状態推移環境における意思決定問題についても多数の研究がある。しかし、現実として重要な問題であるにもかかわらず、両者を統合した問題については限定した取り組みしか行われていなかった。１点目の観点において本実施形態の技術は、状態推定と意思決定問題とを統合を手法である。

【００５０】

２点目の独自の形式の法則とした点について説明する。本実施形態では、観測則と推移則とを、ｐ（s_t'｜ｏｔ）とｐ（a_t｜s_t'，s_{t+1}'）の形式としたことである。この定式化により、確率モデルの上で観測o_tが推定すべき値でなく、すでに与えられた決定事項として扱うことができる。通常の形で定式化した場合、観測o_tは報酬と無関係にすべて別のものとして真の推移則全体を推定しようとする。一方。本実施形態の手法では観測o_tではなく、報酬が予測できれば十分となるため、真の推移則ではなく報酬の予測に必要な隠れ状態s_t'のみにより推移則が再構成される。このことにより、隠れ状態の数が少なくなり、学習を効率良く行うことが可能である。人間の認知などにおいて、視覚及び聴覚などの膨大な観測に対して、意思決定に真に必要な状態は少数である。本法則を用いることにより、このような高度なメカニズムをシンプルなモデルで効果的に実現可能である。

【００５１】

３点目の手順におけるサンプリングの活用について説明する。不確実な内部モデルに対して、現在の推定に基づいてできるだけ報酬を得ようとする活用か、将来の報酬のために情報を得る探索か、のどちらを行うかが重要な問題である。本実施形態では、原理において述べた（２）及び（３）の手順により活用を行うが、元々の推定の不確定性を（１）の

サンプリングで考慮しているので、最も効果的な探索を行うことができる。よって、状態推定モデルとサンプリングとによる意思決定を組み合わせたことによる効率化が図られている点が新規である。

【００５２】

以上に示したように、本発明の実施形態の技術は、演算に係る効率を向上させることを可能とする。すなわち、状態と行動に依存して隠れ状態が推移し、観測状態が生成される隠れマルコフモデルに、報酬が付加されたモデルに対して、機会損失を少なくしつつ、少ない回数で状態の遷移を正しく推定し、結果の説明性が高いアルゴリズムを提供する。

【００５３】

機会損失とは、エージェントの行動に関する損失である。例えば、ａ１という行動を取れば報酬がｒ１得られるにも関わらず、ａ２という行動でｒ２を得た場合、ｒ１－ｒ２が機会損失となる。初期状態で情報が不完全な場合には機会損失を０にすることはできないため、不完全な情報に従って行動したり、情報が十分あるのにも関わらず探索的な行動を取ったりする場合に、機会損失が大きくなるという性質がある。アルゴリズムを長期間繰り返し実行した場合に平均機会損失が少ないことが重要であり、機会損失を少なくする、ことと、トータルの報酬獲得を大きくする、こととはほぼ同じ意味を表す。

【００５４】

［実験結果］

本実施形態の手法の実験結果を説明する。図５は、本実施形態の手法と他の手法の実験結果の一例を示すグラフである。図５は、推定手法ごとの試行回数ごとの累積報酬を示している。本実施形態の提案手法は、O_{mul}→Ｓ→Ｒである。本実施形態の提案手法が最も早く最適な報酬を得られている。また、図６は、実験における収束時の隠れ状態数を表にした図である。本実施形態の手法では、観測状態が３２種類に対して報酬と関係のある８種類の隠れ状態を推定することができている。このことが、高速に学習を行える要因である。

【００５５】

なお、本実施形態の法則に関して補足する。図１で示されている法則は、エージェントには知らされない「真の法則」であり、この法則においては隠れ状態から観測が生成される。通常、確率モデルにおいては真の法則を求めることが一般的であるが、本実施形態で想定している問題設定では必ずしも複雑な真の法則すべてを推定することが必要ではない。特に状態ｓから観測ｏが得られるという部分については，次回どのような観測が得られるかという法則を学習することになるが、実際には状態ｏは観測できるため予測は不要であり、隠れ状態ｓと報酬ｒさえ予測できればよい。よって、ｏをあえて予測しないですむ定式化を考えた結果が本実施形態の手法である。

【００５６】

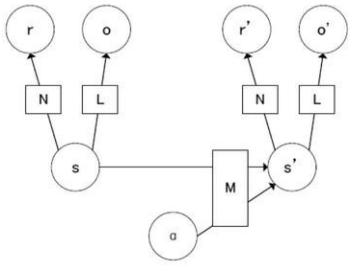
例えば、真の法則において、４つの状態｛ｓ０，ｓ１，ｓ２，ｓ３｝と１対１で観測｛ｏ０，ｏ１，ｏ２，ｏ３｝が対応しているが、報酬の観点からは４つの状態が等価である場合を考えると、報酬を得るという目的のためにはどの観測であっても等価といえる。従来手法では、ｓ０という状態からｏ０が、ｓ１という状態からｏ２が生成されるということを区別して学習し、それに伴って状態数も多く必要となっていた。本実施形態の手法では、ｏ０，ｏ１，ｏ２，ｏ３が観測された場合は共通の状態であるということを学習している。真の法則というのも必ずしも１つの見方に固定されるものでなく、状態は同じで観測が確率的に生成されるという解釈もできる。その意味においてこの例ではｓとｓ'とは完全に等価に対応している。

【００５７】

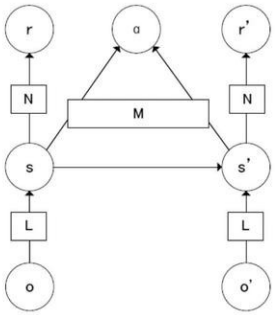
一般に、従来手法のエージェントは必ず観測ｏが説明できるような状態とその推移モデルを構築している。一方、本実施形態の手法では様々な解釈があり得る中で、一見多様で複雑な観測に捉われずに、状態を報酬の観点からシンプルに再構成する。ｐ（ｓ｜ｏ），ｐ（ｏ）からベイズの定理によりｐ（ｏ｜ｓ）が求まることからも，ｓとｓ'とは等価なモデルといえる。

【００５８】
本発明の実施形態の技術は、観測状態と報酬が与えられる広範囲の問題設定に対し適用可能であり、従来の手法に比べて高速に学習可能であるため、多岐にわたる応用が考えられる。特にロボットの分野、及びＡＩを用いたエージェントシステムの高性能化などに活用が期待される。
【００５９】
なお、本発明は、上述した実施形態に限定されるものではなく、この発明の要旨を逸脱しない範囲内で様々な変形や応用が可能である。
【符号の説明】
【００６０】
１００ 最適化装置
１１０ 設定部
１１２ 更新部
１２０ 記憶部

【図１】

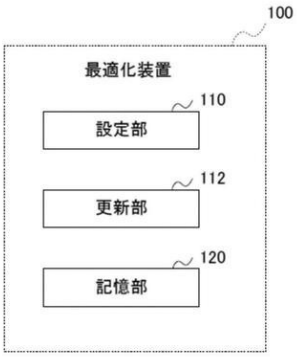


従来の状態及び法則の推定に係る遷移図

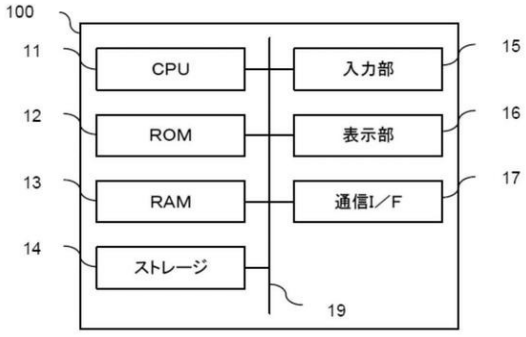


本実施形態の状態及び法則の推定に係る遷移図

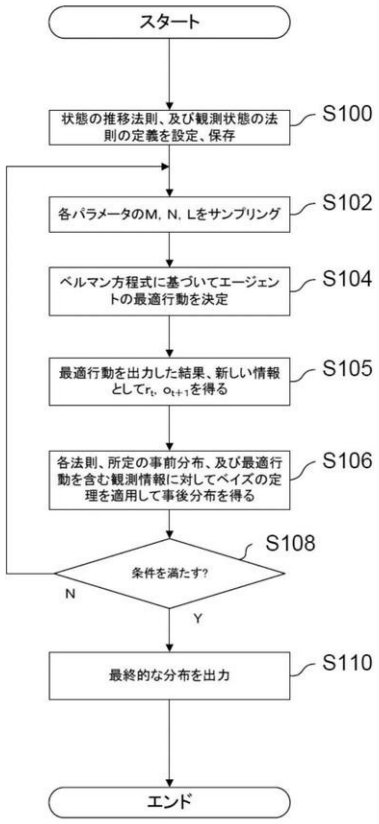
【図２】



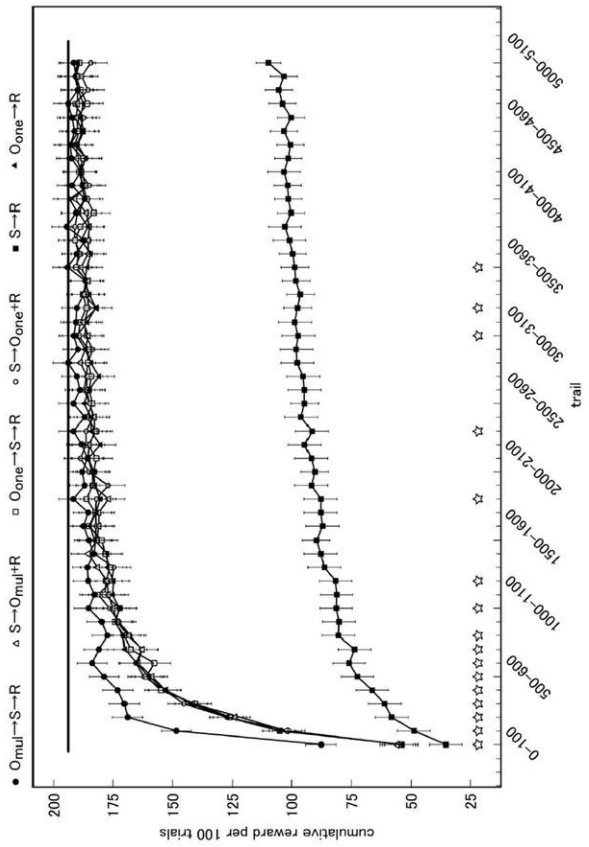
【図３】



【図４】



【図５】



【図６】

モデル	Omul→S→R（本提案手法）	Oone→S→R	S→Omul+R	S→Oone+S
初期値|S|=8	1453	1505	2543	3008
収束時の隠れ状態数	|S|=8	|S|=8	|S|=8	|S|=8
初期値：|S|=32	1461	1492	3102	3078
収束時の隠れ状態数	|S|=8	|S|=32	|S|=32	|S|=32

フロントページの続き

(72)発明者　酒井　裕
東京都新宿区市谷柳町１５－１－１２０４
(72)発明者　深井　朋樹
沖縄県国頭郡恩納村字谷茶１９１９番地１　学校法人沖縄科学技術大学院大学学園内

Ｆターム(参考) 5L049 AA04