(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4811997号
(P4811997)
(45)発行日　平成23年11月9日(2011.11.9)　(24)登録日　平成23年9月2日(2011.9.2)

(51)Int.Cl.　　F I
G05B 13/02 (2006.01)　　G05B 13/02　L
G05B 13/04 (2006.01)　　G05B 13/04

請求項の数 4　(全 17 頁)

(21)出願番号　特願2005-320988 (P2005-320988)
(22)出願日　平成17年11月4日 (2005.11.4)
(65)公開番号　特開2007-128318 (P2007-128318A)
(43)公開日　平成19年5月24日 (2007.5.24)
審査請求日　平成20年8月20日 (2008.8.20)

(73)特許権者　503360115
独立行政法人科学技術振興機構
埼玉県川口市本町四丁目１番８号
(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　100078868
弁理士　河野　登夫
(74)代理人　100114557
弁理士　河野　英仁
(72)発明者　森本　淳
埼玉県川口市本町四丁目１番８号　独立行政法人科学技術振興機構内

最終頁に続く

(54)【発明の名称】状態推定装置、状態推定システム及びコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
観測対象の状態を観測した観測結果に基づいて前記観測対象の状態を推定する状態推定装置において、
前記観測対象の状態を推定する模擬モデルと、
該模擬モデルによる状態の推定結果に基づいて、観測結果を推定した推定観測結果を算出する手段と、
推定観測結果及び観測結果、並びに推定観測結果及び観測結果の差を用いて、強化学習モジュールにおける状態推定の方策に基づくフィードバック値を算出する強化学習モジュールと、
推定観測結果及び観測結果の差並びにフィードバック値に基づいて、報酬値を算出する手段と、
算出した報酬値を用いて強化学習モジュールの方策を更新する更新手段と、
前記強化学習モジュールにて算出されたフィードバック値に基づいて、前記模擬モデルの観測対象の状態を推定する手段と
を備え、
前記更新手段は、報酬値に応じて更新される学習パラメータに基づく平均値及び標準偏差にて示される正規分布に従って分布するように方策を更新する
ことを特徴とする状態推定装置。
【請求項２】

前記学習パラメータは、報酬値の移動平均を用いた関数に基づいて更新される様に構成してあることを特徴とする請求項１に記載の状態推定装置。
【請求項３】
観測対象と、
該観測対象の状態を推定する請求項１又は請求項２に記載の状態推定装置と、
前記観測対象を制御する制御装置と
を備え、
前記状態推定装置の模擬モデルは、前記観測対象の状態の推定結果を前記制御装置へ出力する手段を更に備え、
前記制御装置は、
受け付けた推定結果に基づいて、観測対象を制御する制御命令を生成する手段と、
生成した制御命令を前記観測対象へ出力する手段と
を備え、
前記観測対象は、受け付けた制御命令に従って動作する手段を備える
ことを特徴とする状態推定システム。
【請求項４】
観測対象の状態を観測した観測結果の入力を受け付けるコンピュータに、受け付けた観測結果に基づいて、前記制御対象の状態を推定させるコンピュータプログラムにおいて、
コンピュータに、前記観測対象の状態を模する模擬モデルを用いて、前記観測対象の状態を推定させる手順と、
コンピュータに、前記模擬モデルによる状態の推定結果に基づいて、観測結果を推定した推定観測結果を算出させる手順と、
コンピュータに、推定観測結果及び観測結果、並びに推定観測結果及び観測結果の差を用いて、強化学習モジュールにおける状態推定の方策に基づくフィードバック値を算出させる手順と、
コンピュータに、推定観測結果及び観測結果の差並びにフィードバック値に基づいて、報酬値を算出させる手順と、
コンピュータに、算出した報酬値を用いて強化学習モジュールの方策を更新させる手順と、
コンピュータに、前記強化学習モジュールにて算出されたフィードバック値に基づいて、前記模擬モデルの観測対象の状態を推定させる手順と
を実行させ、
前記更新させる手順は、報酬値に応じて更新される学習パラメータに基づく平均値及び標準偏差にて示される正規分布に従って分布するように方策を更新する
ことを特徴とするコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、観測対象の状態を観測した観測値に基づいて前記観測対象の状態を推定する状態推定装置、該状態推定装置を用いた状態推定システム、及び前記状態推定装置を実現するためのコンピュータプログラムに関し、特に線形又は非線形な動きを示し、その動きを観測することで観測対象の状態を推定する状態推定装置、状態推定システム及びコンピュータプログラムに関する。
【背景技術】
【０００２】
ロボット等の制御対象を制御する場合、制御対象に組み込まれた各種センサがノイズにより、又はセンサそのものを組み込むことができないことにより、必要な状態変数を直接測定することができない状態が生じうる。また例えば市場予測を行う場合においても、価格変動の要因となる株価等の変数は直接観測することができない。
【０００３】

状態変数を直接観測することができない場合、状態観測器（オブザーバ）（非特許文献１参照）、カルマンフィルタ（非特許文献２、３参照）等の方法が一般に用いられている。

【非特許文献１】ディー．ジー．ルエンバーガー（D.G.Luenberger），Anintroduction to observers，IEEETrans.,AC,Vol.16，P.596-602，1971

【非特許文献２】アール．イー．カルマン、アール．エス．ビューシー（R.E.Kalman and R.S.Bucy），New results in linear filtering and prediction theory，Trans.,ASME,Series D,J.ofBasic Engineering,Vol.83,No.1，P.95-108，1961

【非特許文献３】エフ．エル．ルイス（F.L.Lewis），Optimal Estimation:with an Introduction to Stochastic Control Theory，John Wilkey & Sons，1977

【発明の開示】

【発明が解決しようとする課題】

【０００４】

しかしながら非特許文献１の状態観測器及び非特許文献２、３のカルマンフィルタでは、対象のダイナミクスが非線形である場合に、隠れ状態の推定が困難になるという問題がある。

【０００５】

本発明は斯かる事情に鑑みてなされたものであり、観測対象の状態を模する模擬モデル、模擬モデルによる観測対象の状態の推定結果等に基づき状態推定の方策を示すフィードバック値を算出する強化学習モジュール、強化学習モジュールが算出したフィードバック値及び観測対象の観測結果等に基づき報酬値を算出する報酬関数等のプログラムモジュール及び関数を用い、強化学習モジュールでは、フィードバック値が適正な値になる様に報酬値に基づく強化学習を行うことで、線形のダイナミクスだけでなく、非線形のダイナミクスへも容易に適用することが可能な状態推定装置、該状態推定装置を用いた状態推定システム、及び前記状態推定装置を実現するためのコンピュータプログラムの提供を目的とする。

【課題を解決するための手段】

【０００８】

第１発明に係る状態推定装置は、観測対象の状態を観測した観測結果に基づいて前記観測対象の状態を推定する状態推定装置において、前記観測対象の状態を推定する模擬モデルと、該模擬モデルによる状態の推定結果に基づいて、観測結果を推定した推定観測結果を算出する手段と、推定観測結果及び観測結果、並びに推定観測結果及び観測結果の差を用いて、強化学習モジュールにおける状態推定の方策に基づくフィードバック値を算出する強化学習モジュールと、推定観測結果及び観測結果の差並びにフィードバック値に基づいて、報酬値を算出する手段と、算出した報酬値を用いて強化学習モジュールの方策を更新する更新手段と、前記強化学習モジュールにて算出されたフィードバック値に基づいて、前記模擬モデルの観測対象の状態を推定する手段とを備え、前記更新手段は、報酬値に応じて更新される学習パラメータに基づく平均値及び標準偏差にて示される正規分布に従って分布するように方策を更新することを特徴とする。

【０００９】

本発明では、強化学習モジュールが、フィードバック値が適正な値になる様に報酬値に基づく強化学習を行うことで、線形のダイナミクスだけでなく、非線形のダイナミクスへも容易に適用することが可能である。

【００１１】

本発明では、状態推定の方策を示すフィードバック値を、学習パラメータに基づく分布関数に従って様々な方策をとるべく展開させることにより、模擬モデルの推定観測結果が観測対象の観測結果に近付く様に、分布関数の平均値及び標準偏差が更新されながら強化学習が繰り返されるので、観測対象の実際の状態を正確に推定することができる様に、フィードバック値が収束していき、非線形のダイナミクスへも容易に展開することが可能である。

【００１２】
第２発明に係る状態推定装置は、第１発明において、前記学習パラメータは、報酬値の移動平均を用いた関数に基づいて更新される様に構成してあることを特徴とする。
【００１３】
本発明では、強化学習を用いて学習パラメータを更新することにより、ある瞬間での推定誤差を小さくするのではなく、タスクを行っている期間を通じての推定誤差を小さくし、模擬モデルと強化学習モジュールにより推定される状態を適切に観測対象の状態に近付けることが可能である。
【００１４】
第３発明に係る状態推定システムは、観測対象と、該観測対象の状態を推定する第１発明又は第２発明に記載の状態推定装置と、前記観測対象を制御する制御装置とを備え、前記状態推定装置の模擬モデルは、前記観測対象の状態の推定結果を前記制御装置へ出力する手段を更に備え、前記制御装置は、受け付けた推定結果に基づいて、観測対象を制御する制御命令を生成する手段と、生成した制御命令を前記観測対象へ出力する手段とを備え、前記観測対象は、受け付けた制御命令に従って動作する手段を備えることを特徴とする。
【００１５】
本発明では、強化学習モジュールが、フィードバック値が適正な値になる様に報酬値に基づく強化学習を行うことで、線形のダイナミクスだけでなく、非線形のダイナミクスへも容易に適用することが可能であり、しかも強化学習モジュールの学習結果に基づき模擬モデルにて推定された推定結果に基づいて、観測対象を制御することにより、ロボットの制御等の様々な分野への展開が可能である。
【００１６】
第４発明に係るコンピュータプログラムは、観測対象の状態を観測した観測結果の入力を受け付けるコンピュータに、受け付けた観測結果に基づいて、前記制御対象の状態を推定させるコンピュータプログラムにおいて、コンピュータに、前記観測対象の状態を模する模擬モデルを用いて、前記観測対象の状態を推定させる手順と、コンピュータに、前記模擬モデルによる状態の推定結果に基づいて、観測結果を推定した推定観測結果を算出させる手順と、コンピュータに、推定観測結果及び観測結果、並びに推定観測結果及び観測結果の差を用いて、強化学習モジュールにおける状態推定の方策に基づくフィードバック値を算出させる手順と、コンピュータに、推定観測結果及び観測結果の差並びにフィードバック値に基づいて、報酬値を算出させる手順と、コンピュータに、算出した報酬値を用いて強化学習モジュールの方策を更新させる手順と、コンピュータに、前記強化学習モジュールにて算出されたフィードバック値に基づいて、前記模擬モデルの観測対象の状態を推定させる手順とを実行させ、前記更新させる手順は、報酬値に応じて更新される学習パラメータに基づく平均値及び標準偏差にて示される正規分布に従って分布するように方策を更新することを特徴とする。
【００１７】
本発明では、汎用コンピュータ等のコンピュータにて実行することにより、コンピュータが状態推定装置として動作し、強化学習モジュールが、フィードバック値が適正な値になる様に報酬値に基づく強化学習を行うことで、線形のダイナミクスだけでなく、非線形のダイナミクスへも容易に適用することが可能である。
【発明の効果】
【００１８】
本発明に係る状態推定装置、状態推定システム及びコンピュータプログラムは、様々な装置、自然現象、更には経済現象等の線形又は非線形な動きを示し、その状態の一部又は全部を観測することが可能な観測対象と、該観測対象の状態を観測した観測結果に基づいて前記観測対象の状態を推定する状態推定装置とを備え、特に制御可能な様々な装置等の観測対象を観測する場合に、該観測対象を制御する制御装置を更に備える。そして前記観測対象の状態を模する模擬モデルを用いて、前記観測対象の状態を推定し、前記模擬モデ

ルによる状態の推定結果に基づいて、観測結果を推定した推定観測結果を算出し、推定観測結果及び観測結果、並びに推定観測結果及び観測結果の差を用いて推定結果を評価した報酬値に基づいて、状態推定の方策を示すフィードバック値を算出し、推定観測結果及び観測結果の差並びにフィードバック値に基づいて、報酬値を算出し、フィードバック値に基づいて、前記模擬モデルの観測対象の状態推定方法を更新する。

【００１９】

この構成により、本発明では、強化学習モジュールが、フィードバック値が適正な値になる様に報酬値に基づく強化学習を行うことで、線形のダイナミクスだけでなく、非線形のダイナミクスへも容易に適用することが可能である等、優れた効果を奏する。

【００２０】

さらに前記強化学習モジュールは、報酬値に応じて更新される学習パラメータに基づく平均値及び標準偏差にて示される正規分布に従って分布するフィードバック値を算出する様に構成してあり、報酬値の移動平均を用いた関数に基づいて学習パラメータを更新する様に構成してある。

【００２１】

この構成により、本発明では、強化学習に基づいて学習パラメータを更新するため、ある瞬間での推定誤差を小さくするのではなく、タスクを行っている期間を通じての推定誤差を小さくし、模擬モデルと強化学習モジュールにより推定される状態を適切に観測対象の状態に近付けることが可能である等、優れた効果を奏する。

【発明を実施するための最良の形態】

【００２２】

以下、本発明をその実施の形態を示す図面に基づいて詳述する。図１は、本発明の状態推定方法の構成例を概念的に示すブロック図である。図１中１は、観測対象（Ｐｌａｎｔ）であり、本発明の状態推定方法は、観測対象１の状態を推定することを目的とする。観測対象１は、様々な装置、自然現象、更には経済現象等の線形又は非線形な動きを示し、その状態の一部又は全部を観測することが可能である。なお本実施の形態では、観測対象１は、制御装置（Ｃｏｎｔｒｏｌｌｅｒ）２により制御することが可能な装置として以降の説明を行う。また観測対象１の状態は、強化学習状態推定装置（ＲＬＳＥ：Ｒｅｉｎｆｏｒｃｅｍｅｎｔ Ｌｅａｒｎｉｎｇ Ｓｔａｔｅ Ｅｓｔｉｍａｔｏｒ）３により推定される。

【００２３】

強化学習状態推定装置３は、例えば汎用コンピュータにて構成されており、観測対象１の状態を推定する模擬モデル（Ｐｌａｎｔ Ｍｏｄｅｌ）３ａとして機能するモジュールと、模擬モデル３ａに正確な状態を推定させるべく状態推定のための方策を出力する強化学習モジュール（Ｒｅｉｎｆｏｒｃｅｍｅｎｔ Ｌｅａｒｎｉｎｇ Ｍｏｄｕｌｅ）３ｂとを有している。強化学習モジュール３ｂは本発明の状態推定方法において主要な処理を行うプログラムモジュールであり、後述する様々な関数、定数、その他設定値を用いて強化学習を行い、模擬モデル３ａの動きを観測対象１に近付ける方策を展開する。

【００２４】

さらに強化学習状態推定装置３は、模擬モデル３ａが推定した状態から観測結果を推定した推定観測結果を算出する出力関数３ｃ、模擬モデル３ａの推定結果を評価する報酬値を算出する報酬関数３ｄ等の様々な関数を有している。

【００２５】

この様に構成される本発明の状態推定方法において、観測対象１は、制御装置２からの制御命令に基づく制御により動作し、また観測対象１の状態を観測した観測結果は、強化学習状態推定装置３へ入力される。

【００２６】

強化学習状態推定装置３の模擬モデル３ａは、設定されている推定のためのモデルに従って観測対象１の状態を推定し、観測対象１の推定結果として制御装置２へ出力する。なお模擬モデル３ａは、強化学習モジュール３ｂにおける状態推定の方策から出力されたフ

ィードバック値の入力に基づいて、状態の推定を行う。そして制御装置２は、入力を受け付けた推定結果に基づいて観測対象１を制御する制御命令を生成し、生成した制御命令を観測対象１及び強化学習状態推定装置３へ出力する。

【００２７】

なお模擬モデル３ａから出力される推定結果は、強化学習状態推定装置３内の出力関数３ｃ及び強化学習モジュール３ｂでも用いられる。強化学習状態推定装置３では、模擬モデル３ａにより推定した観測対象の状態を示す推定結果から、出力関数３ｃにより、観測対象１の観測結果を推定した推定観測結果を算出し、算出した推定観測結果を報酬関数３ｄへ渡す。

【００２８】

報酬関数３ｄは、観測対象１の観測結果、及び出力関数３ｃにて算出された推定観測結果、並びに強化学習モジュール３ｂにおける状態推定の方策から出力されたフィードバック値に基づいて推定を行い、その結果を評価した報酬値を算出し、算出した報酬値を強化学習モジュール３ｂへ渡す。

【００２９】

強化学習モジュール３ｂは、観測対象１の観測結果及び制御装置２から出力された制御命令、並びに模擬モデル３ａによる推定結果及び報酬関数３ｄにて算出した報酬値に基づいて、状態推定の方策を更新し、更新した方策から出力されるフィードバック値を模擬モデル３ａ及び報酬関数３ｄへ渡す。

【００３０】

図２は、本発明の強化学習状態推定装置３の構成例を示すブロック図である。汎用コンピュータ等のコンピュータを用いた強化学習状態推定装置３は、装置全体を制御するＣＰＵ等の制御手段３１と、本発明の強化学習状態推定装置用のコンピュータプログラム３００及びデータ等の各種情報を記録したＣＤ－ＲＯＭ等の記録媒体３０１から各種情報を読み取るＣＤ－ＲＯＭドライブ等の補助記憶手段３２と、補助記憶手段３２により読み取った各種情報を記録するハードディスク等の記録手段３３と、制御手段３１の制御により一時的に発生するデータを記憶するＲＡＭ等の記憶手段３４とを備えている。そして記録手段３３に記録したコンピュータプログラム３００を記憶手段３４に記憶して制御手段３１の制御により実行することで、汎用コンピュータは、本発明の強化学習状態推定装置３として動作する。さらに強化学習状態推定装置３は、観測対象１の観測結果及び制御装置２の制御命令の入力を受け付ける入力手段３５並びに制御装置２へ推定結果を出力する出力手段３６を備えている。

【００３１】

さらに強化学習状態推定装置３が備える記録手段３３には、模擬モデル３ａ、強化学習モジュール３ｂ、出力関数３ｃ、報酬関数３ｄ等の様々なプログラムモジュール及び関数が記録されている。

【００３２】

次に本発明の強化学習状態推定装置３を用いた状態推定方法について説明する。本発明の状態推定方法は、観測対象１の推定すべき真の状態ｘ及び観測対象１の観測結果（出力値）ｙを用いた下記の式１～式３にて示される。

【００３３】

【数１】

$$\dot{x} = f(x,u) + n(t) \qquad \cdots \text{式}1$$

但し、$\dot{x}$：観測対象１の真の状態の時間微分$(= dx/dt)$

$f(x,u)$：観測対象１のダイナミクス

x：観測対象１の真の状態

u：観測対象１のダイナミクスへの制御入力

$n(t)$：時刻tの関数として示されるノイズ入力

$$y = h(x) + v(t) \qquad \cdots \text{式}2$$

但し、y：観測対象１の観測結果（出力値）

$h(x)$：状態xに基づき観測結果yを決定する出力関数

$v(t)$：時刻tの関数として示される観測ノイズ

$$\dot{\hat{x}} = f(\hat{x},u) + a(\hat{x},y) \qquad \cdots \text{式}3$$

但し、$\dot{\hat{x}}$：観測対象１の推定状態$\hat{x}$の時間微分$(= d\hat{x}/dt)$

$f(\hat{x},u)$：観測対象１の推定状態$\hat{x}$を用いて示されるダイナミクス

$a(\hat{x},y)$：観測対象１の推定状態$\hat{x}$及び観測結果yに基づき決定される状態推定の方策に基づくフィードバック値

【００３４】

なお上記式１において、ノイズ入力ｎ（ｔ）は、下記の式４となる性質を有している。

【００３５】

【数２】

$$E\left[n(t)n^T(\tau)\right] = U\delta(t-\tau) \qquad \cdots \text{式}4$$

但し、$E\left[n(t)n^T(\tau)\right]$：プロセスノイズの共分散行列

Q：プロセスノイズの大きさを示す行列

$\delta()$：ディラックのデルタ関数

τ：時刻（ｔとは異なる）

【００３６】

また上記式２において、観測ノイズｖ（ｔ）は、下記の式５となる性質を有している。

【００３７】

【数３】

$$E\left[v(t)v^{T}(\tau)\right]=S\delta(t-\tau) \qquad \cdots 式5$$

但し、$E\left[v(t)v^{T}(\tau)\right]$：観測ノイズの共分散行列

S：観測ノイズの大きさを示す行列

【００３８】

上記式１において、観測対象１のダイナミクスｆ（ｘ，ｕ）は、観測対象１の観測結果ｙに基づいて、観測対象１の推定状態x^（以降の文章中において、観測対象１の推定状態（推定結果）を示す記号をx^と表記する。）と真の状態ｘとの差異を小さくすべく強化学習することにより取得する関数である。但し、予め既知のダイナミクスｆ（ｘ，ｕ）を設定しておく様にしても良い。なお本発明の状態推定方法では、模擬モデル３ａにより推定した観測対象の状態ｘの推定結果x^から算出した推定観測結果y^（以降の文章中において、推定観測結果を示す記号をy^と表記する。）と、観測対象１の実際の観測結果ｙとの差異が小さい程、高い評価となる下記の式６の報酬関数を用いることにより、推定結果x^を実際の状態ｘに近付けるべく強化学習を行う。

【００３９】

【数４】

$$r(\hat{x},y,a)=e\left(\left|y-h(\hat{x})\right|\right)-c(a) \qquad \cdots 式6$$

但し、$r(\hat{x},y,a)$：推定結果$\hat{x}$、観測結果y及びフィードバックパラメータaに基づく報酬関数３ｄ

$e\left(\left|y-h(\hat{x})\right|\right)$：推定結果の正確さを評価する関数

$h(\hat{x})$：推定状態$\hat{x}$から推定観測結果$\hat{y}$を算出する出力関数３ｃ

$c(a)$：方策の出力ａに対するコスト関数

a：模擬モデル３ａの状態推定の方策に基づくフィードバック値

【００４０】

式６中の推定結果の正確さを評価する関数ｅ（）は、実際の観測結果ｙと、出力関数３ｃとの差異が小さい程、大きい値をとり、関数ｃ（ａ）は、強化学習モジュール３ｂにおける方策の出力の大きさが大きい程、大きい値をとるように設定される。

【００４１】

強化学習モジュール３ｂでは、式６に示す報酬関数ｒに基づく学習処理を繰り返すことで、模擬モデル３ａの状態推定の方策を更新し、更新した状態推定の方策からフィードバック値ａを算出し、模擬モデル３ａへ出力する。なお状態推定の方策を示すフィードバック値ａは、不明な状態から下記の式７に示す確率分布に従って様々な方策をとるべく展開され、模擬モデル３ａの推定観測結果y^が観測対象１の観測結果ｙに近付く様に、強化学習の繰り返しにて収束する。

【００４２】

【数５】

$$\pi_w(a_j \mid \hat{x}, y) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \frac{-(a_j - \mu_j(\hat{x}, y))^2}{2\sigma_j^2} \quad \cdots \text{式７}$$

但し、$\pi_w(\ )$:方策を表す確率分布

j:状態推定の方策の出力回数を示す自然数

μ:状態推定の方策を示すフィードバックパラメータａの平均値

σ^2:分散

【００４３】

式７にて示される様にフィードバック値ａは、平均値μ、分散σ^2の正規分布を示し、分散σ^2が小さい程、方策のバラツキが小さくなる。

【００４４】

式７中の平均値μは、下記の式８～式１０にて示され、学習すべきパラメータを更新していくことにより学習が行われる。

【００４５】

【数６】

$$\mu_j(\hat{x}, y) = \nu_j(\tilde{x})(y - h(\hat{x})) \quad \cdots \text{式８}$$

但し、$\nu_j(\tilde{x})$:下記の式９にて示されるゲイン関数

$\tilde{x}$:下記の式１０にて示される強化学習モジュール３ｂの状態

$$\nu_j(\tilde{x}) = \sum_{i=1}^{N} w_i^{\mu_j} b_i(\tilde{x}) \quad \cdots \text{式９}$$

但し、$b_i(\tilde{x})$:基底関数

$w_i^{\mu_j}$:平均値μに関する学習すべきパラメータ

$$\tilde{x} = \left(\hat{x}^T, (y - h(\hat{x}))^T\right) \quad \cdots \text{式１０}$$

【００４６】

式７中の分散σ^2の平方根である標準偏差σは、下記の式１１にて示され、学習すべきパラメータを更新していくことにより学習が行われる。

【００４７】

【数７】

$$\sigma_j = \frac{1}{1 + \exp(-w_j^{\sigma})} \quad \cdots \text{式１１}$$

但し、w_j^{σ}:標準偏差σに関する学習すべきパラメータ

【００４８】

一方、強化学習においては、式７及び式８のパラメータを更新するために下記の式１２にて示される価値関数Ｑが用いられる。

【００４９】

【数８】

$$Q_w^v(\hat{x},y,a)=\sum_{i,j} v_i^{\mu_j}\phi_i^{\mu_j}(\hat{x},y,a)+\sum_j \upsilon^{\sigma_j}\phi^{\sigma_j}(\hat{x},y,a) \quad \cdots 式１２$$

但し、Q：価値関数

$v_i^{\mu_j}$：価値関数を表現するパラメータ

$\phi_i^{\mu_j}$：下記の式１３にて示される関数

υ^{σ_j}：価値関数を表現するパラメータ

ϕ^{σ_j}：下記の式１４にて示される関数

$$\phi_i^{\mu_j}(\hat{x},y,a)=\frac{\partial \ln \pi_w}{\partial w_i^{\mu_j}}=\frac{a_j-\mu_j}{\sigma_j^2}(y-h(\hat{x}))b_i(\tilde{x}) \quad \cdots 式１３$$

$$\phi^{\sigma_j}(\hat{x},y,a)=\frac{\partial \ln \pi_w}{\partial \omega^{\sigma_j}}=\frac{\left((a_j-\mu_j)^2-\sigma_j^2\right)(1-\sigma_j)}{\sigma_j^2} \quad \cdots 式１４$$

【００５０】

次に学習すべきパラメータの更新方法について説明する。強化学習状態推定装置３の強化学習モジュール３ｂでは、報酬関数３ｄにて算出された報酬値Ｒ（ｔ）に基づき、予め設定されている時定数τを用いた直近の移動平均値を下記の式１５にて計算する。

【００５１】

【数９】

$$\tau\dot{R}(t)=-R(t)+r\left(\hat{x}(t),y(t),a(t)\right) \quad \cdots 式１５$$

但し、τ：移動平均値を算出する期間を示す時定数

r：式６から計算される報酬

【００５２】

また平均報酬の近似誤差Δ（ｔ）を下記の式１６にて計算する。

【００５３】

【数１０】

$$\Delta(t)=\left(r\left(\hat{x}(t),y(t),a(t)\right)-R(t)\right)+\dot{Q}(t) \quad \cdots 式１６$$

【００５４】

式１６の右辺の第１項は、報酬ｒを式１５にて示した報酬値Ｒ（ｔ）の平均値の差であり、右辺の第２項に示した価値関数Ｑ（ｔ）の時間微分（Ｑの上方にドットを付与した記号にパラメータ（ｔ）を付して表記）が大きく近似誤差Δ（ｔ）が大きいほど、現在の状態推定の方策が正しいことになる。更に近似誤差Δ（ｔ）を修正するための価値関数Ｑのパラメータの更新方法を下記の式１７及び式１８に示す。

【００５５】

【数１１】

$$\dot{v}_i^{\mu_j}(t) = \alpha^{\mu}\Delta(t) z_i^{\mu_j}(t) \qquad \cdots 式１７$$

但し、α^{μ}:学習率

z:式１９により算出される値(*eligibility trance*)

$$\dot{v}^{\sigma_j}(t) = \alpha^{\sigma}(t)\Delta(t) z^{\sigma_i}(t) \qquad \cdots 式１８$$

但し、α^{σ}:学習率

z:式２０により算出される値(*eligibility trance*)

【００５６】

また式１７及び式１８に示した関数の出力の時間に関する加重平均値の更新は下記の式１９及び式２０を用いて行われる。

【００５７】

【数１２】

$$\dot{z}_i^{\mu_j}(t) = -\frac{1}{\kappa} z_i^{\mu_j}(t) + \phi_i^{\mu_j}(t) \qquad \cdots 式１９$$

$$\dot{z}^{\sigma_j}(t) = -\frac{1}{\kappa} z^{\sigma_j}(t) + \phi^{\sigma_j}(t) \qquad \cdots 式２０$$

但し、κ:*eligibility trace* の時定数

【００５８】

そして式１３及び式１４にて示した関数の出力を用い下記の式２１及び式２２により学習すべきパラメータを更新する。

【００５９】

【数１３】

$$\dot{w}_i^{\mu_j} = \beta^{\mu} Q\left(\hat{x}(t), y(t), a_j(t)\right) \phi_i^{\mu_j}(t) \qquad \cdots 式２１$$

$$\dot{w}^{\sigma_j} = \beta^{\sigma} Q\left(\hat{x}(t), y(t), a_j(t)\right) \phi^{\sigma_j}(t) \qquad \cdots 式２２$$

但し、β^{μ}:学習率

β^{σ}:学習率

【００６０】

図３は、本発明の強化学習状態推定装置３の処理を示すフローチャートである。上述した様に本発明の強化学習状態推定装置３は、コンピュータプログラム３００を実行する制御手段３１の制御により、強化学習モジュール３ｂにより、上述した様々な関数及び数式にて、状態推定の方策に基づくフィードバック値を算出し（ステップＳ１）、模擬モデル３ａにて式１を用いて観測対象の状態を推定し（ステップＳ２）、模擬モデル３ａによる状態の推定結果に基づいて、出力関数３ｃにより、観測結果を推定した推定観測結果を算出し（ステップＳ３）、式６を用いて推定観測結果及び観測結果の差並びに強化学習モジュール３ｂにて算出されるフィードバック値に基づいて報酬値を算出し（ステップＳ４）、算出した報酬値を用いて強化学習モジュール３ｂの方策を更新する（ステップＳ５）。そして強化学習状態推定装置３は、制御手段３１の制御に基づいて、強化学習モジュール３ｂにより、算出したフィードバック値に基づいて模擬モデル３ａによる観測対象の状態

を推定する（ステップＳ６）。

【実施例】

【００６１】

次に本発明の状態推定方法を用いて行った実験結果について説明する。図４は、本発明の状態推定方法の実験に用いた観測対象である単振り子を示す模式図である。観測対象である単振り子の重りの質量はｍ、振り子長はｌであり、外力Ｔを加えることにより支点ｏを中心とする揺動を開始する。なお図４に示した単振り子の力学モデルは、下記の式２３にて示される。

【００６２】

【数１４】

$$ml^2\dot{\omega} = -\mu\omega - mgl\sin\theta + T \qquad \cdots 式２３$$

但し、m:重りの質量

l:振り子長

ω:角速度

g:重力加速度

θ:鉛直線と振り子とがなす角度

μ:摩擦係数

T:外力

【００６３】

なお当実験において、μ＝０．０１、ｍ＝１．０kg、ｌ＝１．０m、そしてｇ＝９．８m/s^2である。

【００６４】

ここで観測対象である振り子の状態を示す状態ベクトルをｘ＝（θ，ω）Ｔとし、ｕ＝Ｔとすると、上述した式１及び式２は、下記の式２４及び式２５として示される。

【００６５】

【数１５】

$$\dot{x} = \begin{pmatrix} \omega \\ -\frac{g}{l}\sin\theta - \frac{\mu}{ml^2}\omega \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{ml^2} \end{pmatrix} u + n(t) \qquad \cdots 式２４$$

$$y = Cx + v(t) \qquad \cdots 式２５$$

【００６６】

なお式２４及び式２５において、ノイズｎ（ｔ）及びノイズｖ（ｔ）は、上述した式４及び式５に対し、Ｕ＝ｄｉａｇ｛０．０１、０．０１｝、Ｓ＝１．０を代入することにより表現することができる。

【００６７】

また上述した式３の観測対象の推定状態は、当実験において下記の式２６にて示される。

【００６８】

【数１６】

$$\dot{\hat{x}} = \begin{pmatrix} \hat{\omega} \\ -\frac{g}{l}\sin\hat{\theta} - \frac{\mu}{ml^2}\hat{\omega} \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{ml^2} \end{pmatrix} u + a(\hat{x}, y) \qquad \cdots 式２６$$

【００６９】
なお式２６において、右辺の第３項は、状態推定のフィードバック値である。

【００７０】
そして上述した式６に示す報酬関数ｒは、当実験において下記の式２７にて示される。

【００７１】
【数１７】

$$\gamma = \exp\left(-\frac{(y - C\hat{x})^2}{\sigma_\gamma^2}\right) - \sum_{j=1}^{2} c(a_j) \qquad \cdots 式２７$$

【００７２】
なお当実験においてσｒ＝０．５である。また式２７中において、知識の正確さに依存した関数ｃ（ａｊ）は、フィードバック値ａの最大値であるａ^{max}＝（ａ$_1^{max}$，ａ$_2^{max}$）＝（５．０，５．０）を用いた下記の式２８にて示される。

【００７３】
【数１８】

$$c(a_j) = d\int_0^{a_j} s^{-1}\left(\frac{u}{a_j^{max}}\right)du \qquad \cdots 式２８$$

但し、$d = 0.1$

【００７４】
当実験に際し、学習率は、下記の式２９及び式３０にて示される値を用いた。

【００７５】
【数１９】

$$\alpha^\mu = \alpha^\sigma = 2.5\times10^2 \qquad \cdots 式２９$$

$$\beta^\mu = \beta^\sigma = 0.5 \qquad \cdots 式３０$$

【００７６】
さらに当実験に際し、時定数τ＝０．２sec 及びeligibility trace の時定数κ＝０．２sec との条件設定を行った。

【００７７】
また強化学習によるフィードバック値の更新周期は、０．０２sec である。

【００７８】
図５乃至図７は、本発明の状態推定方法の実験結果を示すグラフである。なお図５乃至図６において（ａ）は、角度θのみが観測される場合、（ｂ）は、角速度ωのみが観測される場合、そして（ｃ）は、角度及び角速度ωの線形和θ＋ωのみが観測される場合の結果を示している。図５は、上述した条件に基づく実験において、時間と観測値との関係を示しており、図５（ａ）は、角度θの観測値の経時変化を示しており、図５（ｂ）は、角速度ωの観測値の経時変化を示しており、そして図５（ｃ）は、観測値がθ＋ωである場合の経時変化を示している。図５は、いずれも横軸に時間をとり、縦軸に観測値をとって、その関係を示している。図５に示す様に観測値はいずれも非線形に変動する。

【００７９】
図６は、横軸に時間をとり、縦軸に角度θをとって、実際の角度及び推定した角度の経時変化を示している。図６（ａ）は、図５（ａ）に示した角度の観測値に基づいて推定した角度と、実際の角度との関係を示している。この実験では、図６（ａ）に示す様に１秒程度で推定した角度が実際の角度に一致している。図６（ｂ）は、図５（ｂ）に示した角速度の観測値に基づいて推定した角度と、実際の角度との関係を示している。この実験では、図６（ｂ）に示す様に３秒程度で推定した角度が実際の角度に一致している。図６（ｃ）は、図５（ｃ）に示した角度及び角速度の和の観測値に基づいて推定した角度と、実際の角度との関係を示している。この実験では、図６（ｃ）に示す様に２秒程度で推定した角度が実際の角度に一致している。

【００８０】
図７は、横軸に時間をとり、縦軸に角速度ωをとって、実際の角速度及び推定した角速度の経時変化を示している。図７（ａ）は、図５（ａ）に示した角度の観測値に基づいて推定した角速度と、実際の角速度との関係を示している。この実験では、図７（ａ）に示す様に２秒程度で推定した角速度が実際の角速度に一致している。図７（ｂ）は、図５（ｂ）に示した角速度の観測値に基づいて推定した角速度と、実際の角速度との関係を示している。この実験では、図７（ｂ）に示す様に２秒程度で推定した角速度が実際の角速度に一致している。図７（ｃ）は、図５（ｃ）に示した角度及び角速度の和の観測値に基づいて推定した角速度と、実際の角速度との関係を示している。この実験では、図７（ｃ）に示す様に２秒程度で推定した角度が実際の角度に一致している。

【００８１】
この様に本発明では、非線形のダイナミクスの状態を容易に推定することが可能である。

【００８２】
前記実施の形態では、単振り子を観測対象とする実験を示したが、本発明はこれに限らず、様々な装置、自然現象、更には経済現象等の線形又は非線形な動きを示す観測対象の状態推定に適用することが可能である。

【図面の簡単な説明】

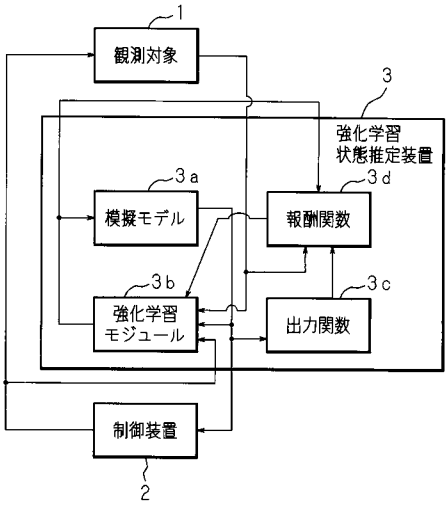
【００８３】
【図１】本発明の状態推定方法の構成例を概念的に示すブロック図である。
【図２】本発明の強化学習状態推定装置の構成例を示すブロック図である。
【図３】本発明の強化学習状態推定装置の処理を示すフローチャートである。
【図４】本発明の状態推定方法の実験に用いた観測対象である単振り子を示す模式図である。
【図５】本発明の状態推定方法の実験結果を示すグラフである。
【図６】本発明の状態推定方法の実験結果を示すグラフである。
【図７】本発明の状態推定方法の実験結果を示すグラフである。

【符号の説明】

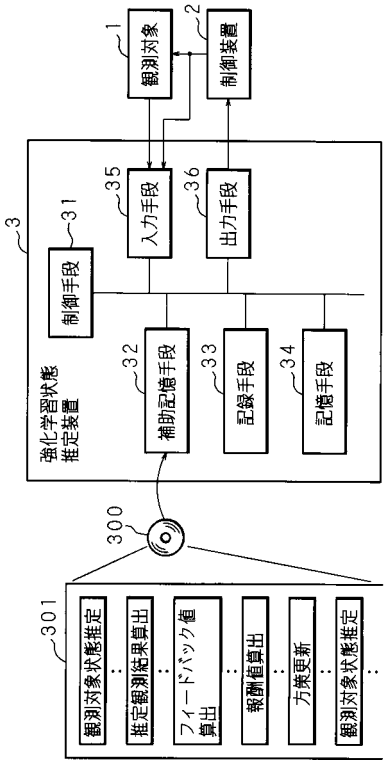
【００８４】
１　観測対象
２　制御装置

３　強化学習状態推定装置
３ａ　模擬モデル
３ｂ　強化学習モジュール
３ｃ　出力関数
３ｄ　報酬関数
３００　コンピュータプログラム
３０１　記録媒体

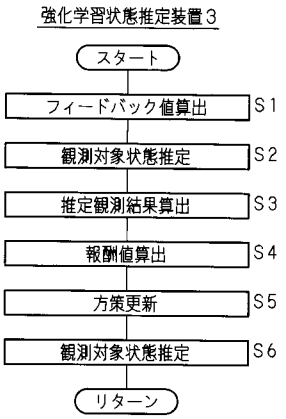
【図１】



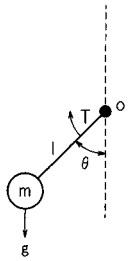
【図２】



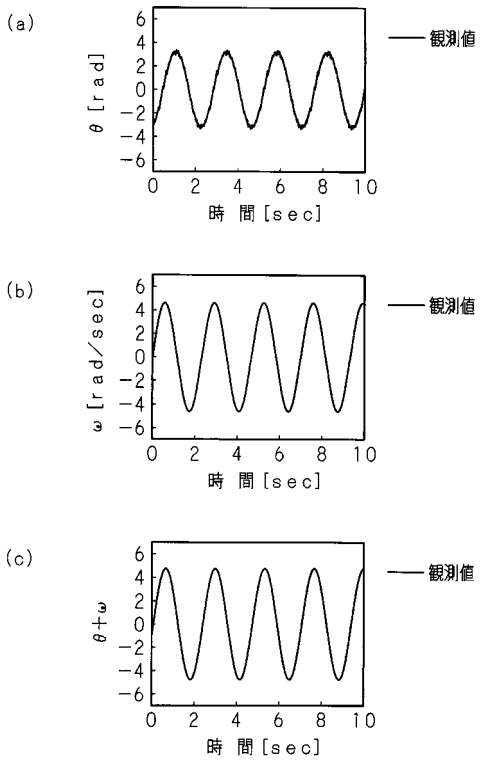
【図３】


強化学習状態推定装置３
スタート
フィードバック値算出
S1
観測対象状態推定
S2
推定観測結果算出
S3
報酬値算出
S4
方策更新
S5
観測対象状態推定
S6
リターン

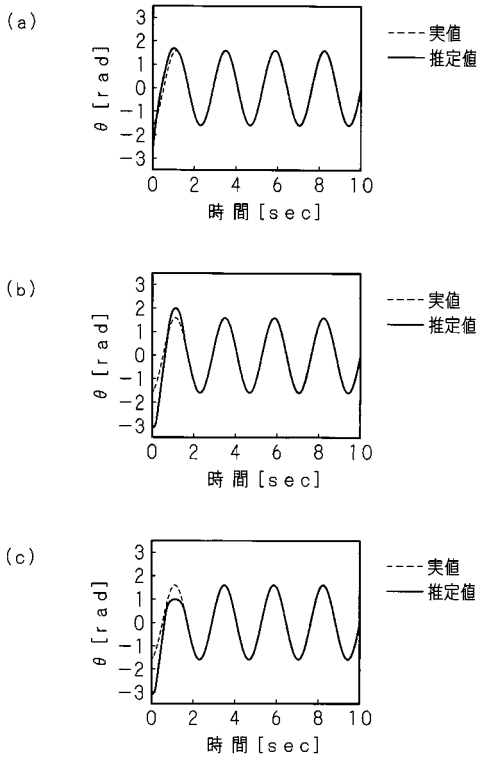
【図４】


T
o
l
θ
m
g

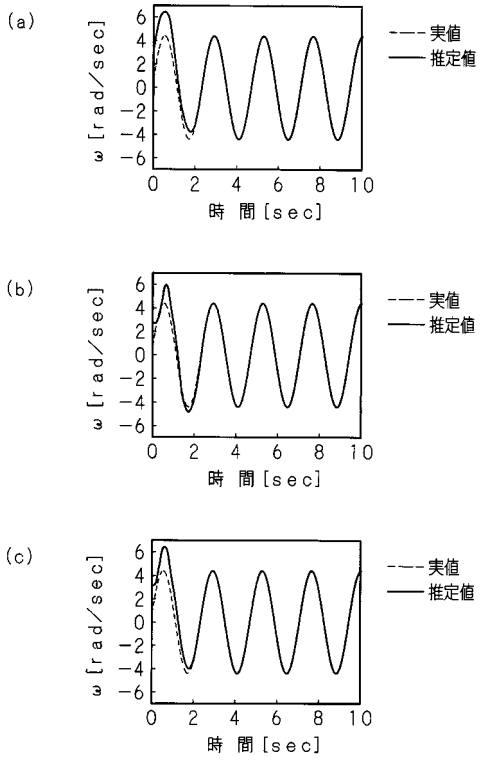
【図５】


(a)
θ [rad]
時 間 [sec]
観測値
(b)
ω [rad/sec]
時 間 [sec]
観測値
(c)
θ+ω
時 間 [sec]
観測値

【図６】


(a)
θ [rad]
時 間 [sec]
実値
推定値
(b)
θ [rad]
時 間 [sec]
実値
推定値
(c)
θ [rad]
時 間 [sec]
実値
推定値

【図７】


(a)
ω [rad/sec]
時 間 [sec]
実値
推定値
(b)
ω [rad/sec]
時 間 [sec]
実値
推定値
(c)
ω [rad/sec]
時 間 [sec]
実値
推定値

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

審査官　星名　真幸

(56)参考文献　森本　淳　Morimoto Jun，銅谷　賢治　Doya Kenji，強化学習を用いた状態観測器の構築　Development of An Observer by using Reinforcement Learning，日本神経回路学会第１２回全国大会講演論文集，日本，日本神経回路学会，２００２年　9月１９日，pp.275-278

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０５Ｂ　１３／０２
Ｇ０５Ｂ　１３／０４