(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5737890号
(P5737890)
(45)発行日　平成27年6月17日(2015.6.17)　(24)登録日　平成27年5月1日(2015.5.1)

(51)Int.Cl.		F I		
G06N 99/00	(2010.01)	G06N	99/00	153
G06N 7/00	(2006.01)	G06N	7/00	150
G06Q 50/04	(2012.01)	G06Q	50/04	
G05B 13/02	(2006.01)	G05B	13/02	L

請求項の数 29　(全 22 頁)

(21)出願番号　特願2010-202137(P2010-202137)
(22)出願日　平成22年9月9日(2010.9.9)
(65)公開番号　特開2011-60290(P2011-60290A)
(43)公開日　平成23年3月24日(2011.3.24)
審査請求日　平成25年9月9日(2013.9.9)
(31)優先権主張番号　10 2009 040 770.7
(32)優先日　平成21年9月9日(2009.9.9)
(33)優先権主張国　ドイツ(DE)

(73)特許権者　390039413
シーメンス　アクチエンゲゼルシヤフト
Ｓｉｅｍｅｎｓ　Ａｋｔｉｅｎｇｅｓｅｌｌｓｃｈａｆｔ
ドイツ連邦共和国　Ｄ－８０３３３　ミュンヘン　ヴィッテルスバッハープラッツ　２
Ｗｉｔｔｅｌｓｂａｃｈｅｒｐｌａｔｚ　２，　Ｄ－８０３３３　Ｍｕｅｎｃｈｅｎ，　Ｇｅｒｍａｎｙ
(74)代理人　100099483
弁理士　久野　琢也
(74)代理人　100061815
弁理士　矢野　敏雄

最終頁に続く

(54)【発明の名称】技術システムの制御および／または調整をコンピュータ支援により学習する方法

(57)【特許請求の範囲】
【請求項１】
技術システムの制御および／または調整をコンピュータ支援により学習する方法であって、
技術システムの運転が、運転中の技術システムの状態（ｓ）と、技術システムの運転中に実行され、技術システムのそれぞれの状態（ｓ）を連続状態に移行させる活動とによって特徴付けられる方法において、
・技術システムの運転中に求められた、状態（ｓ）、活動（ａ）および連続状態（ｓ’）を含むトレーニングデータに基づいて、品質関数（Ｑ）と活動選択ルール（π（ｓ））を学習するステップ；
ただし前記品質関数（Ｑ）は技術システムの最適運転をモデル化し、
前記活動選択ルール（π（ｓ））は、技術システムの運転中に当該技術システムのそれぞれの状態（ｓ）に対して実行すべき活動（ａ）を指示し、
・品質関数（Ｑ）および活動選択ルール（π（ｓ））の学習中に、品質関数（Ｑ）の統計的不確定性に対する尺度（σＱ）を、不確定性伝播によって求めるステップ；
・該統計的不確定性に対する尺度（σＱ）と、品質関数（Ｑ）への統計的な要求に相当する安全パラメータ（ζ）とに基づいて、変形された品質関数を決定するステップ；
ただし前記不確定性伝播は、非対角要素が無視された共分散マトリクスを使用し、
・変形された品質関数に基づいて、活動選択ルール（π（ｓ））を学習するステップ；
を有する方法。

【請求項２】
前記品質関数（Ｑ）を、評価（Ｒ）および状態活動確率（Ｐ）を考慮して学習し、
それぞれの評価（Ｒ）は、状態（ｓ）、当該状態で実行された活動（ａ）、および連続状態（ｓ’）からなる組合せの品質を、技術システムの最適運転の観点で評価し、
それぞれの状態活動確率（Ｐ）は、状態と当該状態で実行された活動（ａ）に依存して、連続状態（ｓ’）の確率（Ｐ）を指示する請求項１記載の方法。
【請求項３】
品質関数（Ｑ）と活動選択ルール（π（ｓ））を、ベルマン反復式に基づいて学習し、
各反復ステップで新たな品質関数（Ｑ）と、該品質関数（Ｑ）の統計的不確定性に対する新たな尺度を求め、それにより新たに変形された品質関数を決定し、
それぞれの反復ステップで共分散マトリクスを、前記品質関数（Ｑ）、状態活動確率（Ｐ）および評価（Ｒ）に依存し、非対角要素を無視して求める請求項２記載の方法。
【請求項４】
ベルマン反復法のｍ番目の反復ステップで、活動選択ルールを以下の活動$\alpha_{s,max}$に基づいて求め、
【数１】

$$\forall s : a_{s,\max} = \arg\max_{a}[Q^m(s,a) - \xi\sigma Q^m(s,a)]$$

ここで
【数２】

$$Q^m(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{m-1}(s')]$$

は品質関数であり、
【数３】

$$Q^m(s,a) - \xi\sigma Q^m(s,a)$$

は変形された品質関数であり、
$\sigma Q^m(s,a)$はm番目の反復ステップにおける品質関数（Ｑ）の統計的不確定性に対する尺度（σＱ）であり、
ここで
【数４】

$$\begin{aligned}(\sigma Q^m(s,a))^2 = &\sum_{s'} (D_{QQ})^2 (\sigma V^{m-1}(s'))^2 + \\ &\sum_{s'} (D_{QP})^2 (\sigma P(s'|s,a))^2 + \\ &\sum_{s'} (D_{QR})^2 (\sigma R(s,a,s'))^2,\end{aligned}$$

$$(D_{QQ}) = \gamma P(s'|s,a),$$
$$(D_{QP}) = R(s,a,s') + \gamma V^{m-1}(s'),$$
$$(D_{QR}) = P(s'|s,a)$$

γ∈［０，１］は非連続因子であり、
ζは安全パラメータであり、

【数５】

$$V^m(s) = \max_a [Q^m(s,a) - \xi\sigma Q^m(s,a)] ;$$

$$(\sigma V^m(s))^2 = (\sigma Q(s, a_{s,\max}))^2 ;$$

が成り立ち、

Ｐ（ｓ’｜ｓ，ａ）は、状態ｓで活動ａが実行された際の連続状態ｓ’に対する状態活動確率であり、

Ｒ（ｓ，ａ，ｓ’）は、状態ｓで活動ａが実行された際の連続状態ｓ’の評価であり、

σＰ（ｓ’｜ｓ，ａ）は、状態－活動確率の統計的不確定性であり、

σＲ（ｓ，ａ，ｓ’）は、評価の統計的不確定性である請求項３記載の方法。

【請求項５】

状態活動確率（Ｐ）を状態活動確率分布としてモデル化し、および／または評価（Ｒ）を評価確率分布としてモデル化する請求項２から４までのいずれか一項記載の方法。

【請求項６】

状態活動確率（Ｐ）の統計的不確定性（σＰ）を、モデル化した状態活動確率分布から求め、評価の統計的不確定性（σＲ）を、モデル化した評価確率分布から求める請求項４および５記載の方法。

【請求項７】

状態活動確率分布および／または評価確率分布を、トレーニングデータからの相対的頻度としてモデル化し、

ここで状態活動確率分布は多項分布としてモデル化し、および／または評価確率分布は正規分布としてモデル化する請求項５または６記載の方法。

【請求項８】

状態活動確率分布を、アプリオリ分布とアポステリオリパラメータを用いたベイズの推定に基づいてモデル化し、ここでアポステリオリパラメータはトレーニングデータに依存する請求項５から７までのいずれか一項記載の方法。

【請求項９】

アプリオリ分布は、ディリクレ分布および／または正規分布である請求項８記載の方法。

【請求項１０】

ディリクレ分布のパラメータ（α_{ijk}）は、連続状態（ｓ’）の平均数と、トレーニングデータによる状態（ｓ）の総数の商に相当する請求項９記載の方法。

【請求項１１】

学習すべき活動選択ルールは、決定論的活動選択ルールである請求項１から１０までのいずれか一項記載の方法。

【請求項１２】

ベルマン反復式のｍ番目の反復ステップにおける活動選択ルールπ^m（ｓ）は以下のとおりであり、

【数６】

$$\pi^m(s) = \arg\max_a Q^m(s,a) - \xi\sigma Q^m(s,a)$$

ここで

【数７】

$$\pi^m(s)$$

は、選択された活動である、請求項４にかかる請求項１１記載の方法。

【請求項１３】

学習すべき活動選択ルールは、技術システムの状態（ｓ）のために実行可能な活動（ａ

）に対する確率分布を指示する確率論的活動選択ルール（π（ｓ））である請求項１から１０までのいずれか一項記載の方法。
【請求項１４】
前記ベルマン反復式の各反復ステップにおいて、実行可能な活動（ａ）に対する新たな確率分布として確率分布を求め、
該確率分布は、最後の反復ステップの確率分布を、変形された品質関数の値を最大にする活動（ａ）に比較的高い確率が割り当てられるよう変形する、請求項３にかかる請求項１３記載の方法。
【請求項１５】
当該方法により、タービンの制御および／または調整が学習される請求項１から１４までのいずれか一項記載の方法。
【請求項１６】
当該方法により、風力発電設備の制御および／または調整が学習される請求項１から１４までのいずれか一項記載の方法。
【請求項１７】
技術システムの運転方法であって、
該技術システムが、請求項１から１６までのいずれか１項記載の方法により学習された制御および／または調整に基づいて運転され、学習された活動選択ルールにより技術システムのそれぞれの状態（ｓ）で実行すべき活動（ａ）が選択される運転方法。
【請求項１８】
技術システムの運転中に、請求項１から１６までのいずれか１項記載の方法が反復され、
各反復の際に、技術システムが取る新たな状態（ｓ）および／または実行すべき活動（ａ）がトレーニングデータとして考慮される請求項１７記載の方法。
【請求項１９】
コンピュータに、
・技術システムの運転中に求められた、状態（ｓ）、活動（ａ）および連続状態（ｓ’）を含むトレーニングデータに基づいて、品質関数（Ｑ）と活動選択ルール（π（ｓ））を学習するステップ；
ただし前記品質関数（Ｑ）は技術システムの最適運転をモデル化し、
前記活動選択ルール（π（ｓ））は、技術システムの運転中に当該技術システムのそれぞれの状態（ｓ）に対して実行すべき活動（ａ）を指示し、
・品質関数（Ｑ）および活動選択ルール（π（ｓ））の学習中に、品質関数（Ｑ）の統計的不確定性に対する尺度（σＱ）を、不確定性伝播によって求めるステップ；
・該統計的不確定性に対する尺度（σＱ）と、品質関数（Ｑ）への統計的な要求に相当する安全パラメータ（ζ）とに基づいて、変形された品質関数を決定するステップ；
ただし前記不確定性伝播は、非対角要素が無視された共分散マトリクスを使用し、
・変形された品質関数に基づいて、活動選択ルール（π（ｓ））を学習するステップ；
を実行させるためのコンピュータプログラム。
【請求項２０】
コンピュータを備えた、技術システムの制御装置であって、
技術システムの運転が、運転中の技術システムの状態（ｓ）と、技術システムの運転中に実行され、技術システムのそれぞれの状態（ｓ）を連続状態に移行させる活動とによって特徴付けられる、制御装置において、
前記コンピュータは、
・技術システムの運転中に求められた、状態（ｓ）、活動（ａ）および連続状態（ｓ’）を含むトレーニングデータに基づいて、品質関数（Ｑ）と活動選択ルール（π（ｓ））を学習し、
ただし前記品質関数（Ｑ）は技術システムの最適運転をモデル化し、
前記活動選択ルール（π（ｓ））は、技術システムの運転中に当該技術システムのそれ

ぞれの状態（ｓ）に対して実行すべき活動（ａ）を指示し、
・品質関数（Ｑ）および活動選択ルール（π（ｓ））の学習中に、品質関数（Ｑ）の統計的不確定性に対する尺度（σＱ）を、不確定性伝播によって求め、
・該統計的不確定性に対する尺度（σＱ）と、品質関数（Ｑ）への統計的な要求に相当する安全パラメータ（ζ）とに基づいて、変形された品質関数を決定し、
ただし前記不確定性伝播は、非対角要素が無視された共分散マトリクスを使用し、
・変形された品質関数に基づいて、活動選択ルール（π（ｓ））を学習する
ことを特徴とする制御装置。
【請求項２１】
前記コンピュータは、
前記品質関数（Ｑ）を、評価（Ｒ）および状態活動確率（Ｐ）を考慮して学習し、
それぞれの評価（Ｒ）は、状態（ｓ）、当該状態で実行された活動（ａ）、および連続状態（ｓ’）からなる組合せの品質を、技術システムの最適運転の観点で評価し、
それぞれの状態活動確率（Ｐ）は、状態と当該状態で実行された活動（ａ）に依存して、連続状態（ｓ’）の確率（Ｐ）を指示する請求項２０記載の制御装置。
【請求項２２】
前記コンピュータは、
品質関数（Ｑ）と活動選択ルール（π（ｓ））を、ベルマン反復式に基づいて学習し、
各反復ステップで新たな品質関数（Ｑ）と、該品質関数（Ｑ）の統計的不確定性に対する新たな尺度を求め、それにより新たに変形された品質関数を決定し、
それぞれの反復ステップで共分散マトリクスを、前記品質関数（Ｑ）、状態活動確率（Ｐ）および評価（Ｒ）に依存し、非対角要素を無視して求める請求項２１記載の制御装置。
【請求項２３】
前記コンピュータは、状態活動確率（Ｐ）を状態活動確率分布としてモデル化し、および／または評価（Ｒ）を評価確率分布としてモデル化する請求項２１または２２記載の制御装置。
【請求項２４】
前記コンピュータは、
状態活動確率（Ｐ）の統計的不確定性（σＰ）を、モデル化した状態活動確率分布から求め、評価の統計的不確定性（σＲ）を、モデル化した評価確率分布から求める請求項２３記載の制御装置。
【請求項２５】
前記コンピュータは、
状態活動確率分布および／または評価確率分布を、トレーニングデータからの相対的頻度としてモデル化し、
ここで状態活動確率分布は多項分布としてモデル化し、および／または評価確率分布は正規分布としてモデル化する請求項２３または２４記載の制御装置。
【請求項２６】
前記コンピュータは、
状態活動確率分布を、アプリオリ分布とアポステリオリパラメータを用いたベイズの推定に基づいてモデル化し、ここでアポステリオリパラメータはトレーニングデータに依存する請求項２３から２５までのいずれか１項記載の制御装置。
【請求項２７】
学習すべき活動選択ルールは、決定論的活動選択ルールである請求項２０から２６までのいずれか１項記載の制御装置。
【請求項２８】
学習すべき活動選択ルールは、技術システムの状態（ｓ）のために実行可能な活動（ａ）に対する確率分布を指示する確率論的活動選択ルール（π（ｓ））である請求項２０から２６までのいずれか１項記載の制御装置。

【請求項２９】
タービンまたは風力発電設備を制御する、
請求項２０から２８までのいずれか１項記載の制御装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、技術システムの閉ループ制御（制御）および／または開ループ制御（調整）をコンピュータ支援により学習する方法、対応する技術システムの運転方法、およびコンピュータプログラムに関する。
【背景技術】
【０００２】
従来技術から、前もって求めたトレーニングデータに基づき（このトレーニングデータは技術システムの運転を表す）、当該システムの最適運転をモデル化することのできる種々の方法が公知である。技術システムは状態、活動および連続状態により記述される。ここで状態とは、特定の技術パラメータまたは技術システムの観察された状態量である。また活動は、対応する調整量を表し、この調整量は技術システムにおいて変化し得る。従来技術から一般的に強化学習方法（英語：Reinforcement Learning）が公知である。この強化学習方法は、技術システムのためにトレーニングデータに基づき、最適化基準にしたがって最適の活動選択ルールを学習する。公知の方法は、学習した活動選択ルールのランダムな不確定性に関しては予測を提供しないという欠点を有する。このような不確定性は、とりわけトレーニングデータ量が小さい場合に非常に大きくなる。
【０００３】
非特許文献１には、活動選択ルールの学習のために使用される品質関数における統計的不確定性を考慮する方法が記載されている。ここでは、活動選択ルールを決定する学習方法が統計的不確定性と組み合わされ、ガウスの誤差伝搬とも称されるそれ自体公知の不確定性伝播（英語：Uncertainty Propagation）に基づいて、学習の際に考慮される品質関数の統計的不確定性が求められる。不確定性伝播では、学習方法で導入される変数の不確定性間の相関が共分散マトリクスによって考慮される。このようにして不確定性は変数内に正確に伝播され、計算される。このことは、技術システムの対応する制御をコンピュータ支援により学習する際には非常に大きな計算コストとメモリスペースを必要とする。
【先行技術文献】
【非特許文献】
【０００４】
【非特許文献１】D. Schneegass, S. Udluft, T. Martinetz: Uncertainty Propagation for Quality Assurance in Reinforcement Learning, 2008, Proc. of the International Joint Conference on Neural Networks (IJCNN), pages 2589-2596.
【発明の概要】
【発明が解決しようとする課題】
【０００５】
したがって本発明の課題は、学習の際に使用されるトレーニングデータの統計的不確定性を考慮し、同時にメモリスペース需要と計算時間に関して効率的な、技術システムの閉ループ制御および／または開ループ制御の学習方法を提供することである。
【課題を解決するための手段】
【０００６】
この課題は独立請求項により解決される。本発明の有利な実施形態は従属請求項に記載されている。
【０００７】
本発明の方法によれば、技術システムの閉ループ制御または開ループ制御がコンピュータ支援により学習される。技術システムの運転は、運転中の技術システムの状態と、技術システムの運転中に実行され、技術システムのそれぞれの状態を連続状態に移行させる活

動とによって特徴付けられる。本発明の方法では、技術システムの運転中に求められたトレーニングデータを含む状態、活動、および連続状態に基づいて、品質関数と活動選択ルールが学習される。ここで学習はとりわけ強化学習方法により行われる。品質関数は、技術システムの最適運転を、この技術システムに対する固有の基準に関してモデル化し、活動選択ルールは技術システムの運転中に、技術システムのそれぞれの状態に対して優先的に実行すべき活動を指示する。

【０００８】

本発明の方法では、品質関数および活動選択ルールの学習中に品質関数の統計的不確定性に対する尺度が、不確定性伝播によって求められ、この統計的不確定性に対する尺度と、品質関数への統計的に緩和された要求に相当する安全パラメータとに基づいて、モデル化された品質関数が決定される。統計的不確定性に対する尺度とは、統計的分散または標準偏差に対する尺度であると理解すべきであり、好ましくは統計的分散または標準偏差自体である。これらから決定されたモデル化された品質関数に基づいて、活動選択ルールが学習される。

【０００９】

非特許文献１の方法との相違は、本発明による方法では、不確定性伝播が共分散マトリクスを使用し、この共分散マトリクスでは非対角要素が無視される、すなわち非対角要素がゼロにセットされることである。したがってこのことは、不確定性伝播の際に考慮される変数間の相関が無視されることと同義である。したがって不確定性はもはや正確には伝播および計算されず、単に近似が実行される。しかしこの近似にもかかわらず、本発明の方法は、確定性が最適である活動選択ルールの形で良好な結果をもたらし、この活動選択ルールは、技術システムのパフォーマンスを統計的不確定性を考慮して最大にする。この方法は、特許文献１の方法に対して、計算時間と必要なワークメモリが格段に小さいという利点を有する。なぜなら共分散マトリクスの対角要素だけを求めれば良いからである。とりわけ計算時間と必要ワークメモリは、統計的不確定性を考慮しない従来の強化学習法と同じオーダーである。

【００１０】

品質関数と活動選択ルールの学習は、本発明の方法の好ましい変形実施形態では、評価と状態活動確率を考慮して行われる。ここでそれぞれの評価では、状態と、この状態で実行される活動と、連続状態との組合せの品質が、技術システムの最適運転の観点で評価される。この評価はしばしば報酬とも称される。状態活動確率は、状態およびこの状態で実行される活動の関数として連続状態の確率を表す。評価が学習の際に考慮されるなら、このような評価はトレーニングデータに含まれるか、または状態、活動および連続状態に依存して相当の評価を送出する関数が抽出される。

【００１１】

とりわけ好ましい実施形態では、品質関数と活動選択ルールが、それ自体公知のベルマン反復法に基づいて学習される。ここで各反復ステップでは、新たな品質関数、品質関数の統計的不確定性に対する新たな尺度、およびこれにより新たに変形された品質関数が決定され、それぞれの反復ステップでは、統計的不確定性に対する新たな尺度を決定するために共分散マトリクスが、品質関数、状態活動確率、および非対角要素を無視した評価によって求められる。したがって分散だけが不確定性伝播に入り込む。すなわち、共分散マトリクスは、品質関数の統計的不確定性と、評価の統計的不確定性と、状態活動確率の統計的不確定性との相関が無視されるようにして近似される。

【００１２】

好ましい実施形態では、ベルマン反復法のｍ番目の反復ステップで、活動選択ルールが以下の活動$\alpha_{s,max}$に基づいて求められる。

【００１３】

【数１】

$$\forall s : a_{s,\max} = \arg\max_{a}[Q^m(s,a) - \xi\sigma Q^m(s,a)]$$

ここで
【数２】

$$Q^m(s,a)=\sum_{s'}P(s'|s,a)[R(s,a,s')+\gamma V^{m-1}(s')]$$

品質関数、そして
【数３】

$$Q^m(s,a)-\xi\sigma Q^m(s,a)$$

変形された品質関数であり、
【数４】

$$\sigma Q^m(s,a)$$

はｍ次の反復ステップでの品質関数（Ｑ)の統計的不確定性に対する尺度（σＱ)である。ただし
【数５】

$$\begin{aligned}(\sigma Q^m(s,a))^2 &= \sum_{s'}(D_{QQ})^2(\sigma V^{m-1}(s'))^2 + \\ &\sum_{s'}(D_{QP})^2(\sigma P(s'|s,a))^2 + \\ &\sum_{s'}(D_{QR})^2(\sigma R(s,a,s'))^2,\end{aligned}$$

$$(D_{QQ})=\gamma P(s'|s,a),$$

$$(D_{QP})=R(s,a,s')+\gamma V^{m-1}(s'),$$

$$(D_{QR})=P(s'|s,a)$$

γ∈［０，１］は不連続関数であり、
ξは安全パラメータであり、
【数６】

$$V^m(s)=\max_a[Q^m(s,a)-\xi\sigma Q^m(s,a)]\ ;$$

$$(\sigma V^m(s))^2=(\sigma Q(s,a_{s,\max}))^2\ \ ;$$

が当てはまる。
Ｐ（ｓ’｜ｓ，ａ）は、状態ｓで活動ａが実行された際の連続状態ｓ’に対する状態活動確率であり、
Ｒ（ｓ，ａ，ｓ’）は、状態ｓで活動ａが実行された際の連続状態ｓ’の評価であり、
σＰ（ｓ’｜ｓ，ａ）は、状態活動確率の統計的不確定性であり、
σＲ（ｓ，ａ，ｓ’）は、評価の統計的不確定性である。
【００１４】
本発明の方法の別のとくに好ましい実施形態では、状態活動確率が状態活動確率分布としてモデル化され、および／または評価が評価確率分布としてモデル化される。状態活動確率分布または評価確率分布は、状態活動確率または評価の統計的不確定性が入り込む上記の方法で好ましくは、この統計的不確定性を決定するために使用される。
【００１５】
本発明の別の変形実施形態では、状態活動確率分布および／または評価確率分布が、トレーニングデータから相対的頻度としてモデル化される。ここで状態活動確率分布はとり

わけ多項分布としてモデル化され、および／または評価確率分布はとりわけ正規分布としてモデル化される。

【００１６】

本発明の方法の別のとくに好ましい実施形態では、状態－活動確率分布が、アプリオリ分布とアポステリオリパラメータを用いたベイズの推定に基づいてモデル化される。ここでアポステリオリパラメータはトレーニングデータに依存する。

【００１７】

このベイズのモデル化は、推定者の不確定性に良好にアクセスできるという利点を有する。ここで好ましくは、アプリオリ分布としてディリクレ分布または場合により正規分布を使用する。とくに好ましい変形実施形態では、ディリクレ分布を使用する場合、その各パラメータが、連続状態の平均数とトレーニングデータによる状態の総数との商に相当するように選択される。このようにして、観察が少数である場合に対しても現実的な活動選択ルールが学習される。

【００１８】

本発明の方法で学習された活動選択ルールは、決定論的であっても確率論的であっても良い。ここで、決定論的活動選択ルールは技術システムの状態に対して明確な活動を指示する。これに対して確率論的活動選択ルールは、技術システムの状態に対して、実行可能な活動に対する確率分布を指示する。欠点論的活動選択ルールπが使用される場合、上記のベルマン反復法のｍ番目の反復ステップにおける活動選択ルールπ^m（ｓ）は次のとおりである。

【００１９】

【数７】

$$\pi^m(s) = \arg\max_a Q^m(s,a) - \xi\sigma Q^m(s,a)$$

ここでπ^m（ｓ）は選択された活動である。

【００２０】

別のとくに好ましい実施形態では確率論的活動選択ルールが次のように構成されている。すなわち、上記のベルマン反復式の各反復ステップにおいて、実行可能な活動に対する新たな確率分布として確率分布が求められ、この確率分布は、最後の反復ステップの確率分布を、変形された品質関数の値を最大にする活動に比較的高い確率が割り当てられるように変形する。

【００２１】

本発明による方法は任意の技術システムに使用可能である。とくに好ましい変形実施形態では、タービン、とりわけガスタービンの制御または調整を学習するための本発明が使用される。ここでガスタービンの状態は、例えば供給される燃料量および／またはタービンのうなりである。活動は例えば、供給される燃料量の変化またはタービンの翼における調整変化である。

【００２２】

本発明の方法の別の変形実施形態では、風力発電設備の制御および／または調整が学習される。ここで風力発電設備の状態は、例えば風力、ロータ回転数、設備のコンポーネントの磨耗等とすることができる。活動はこの関連で、例えば風力発電設備の個々のロータブレードの調整角の調整とすることができる。

【００２３】

上記の学習方法の他に、本発明はさらに技術システムの運転方法を含む。ここで技術システムは、上記の学習方法の任意の変形実施形態により学習された制御または調整に基づいて運転される。技術システムのそれぞれの状態で学習された活動選択ルールにより、実行すべき活動が選択される。確率論的活動選択ルールでは、この選択がそれぞれの確率にしたがった、活動のランダムな選択により行われる。運転の好ましい変形実施形態では、上記の学習方法が間隔を置いて繰り返される。ここでは各繰り返しの際に、技術システムから取り出された新たな状態と実行された活動がトレーニングデータとして考慮される。

【００２４】
上記の方法の他、本発明はさらに、コンピュータに手順を実行させるためのプログラムコードを有するコンピュータプログラムを記録したコンピュータ読み取り可能媒体に関しており、ここでこのプログラムコードは、相応するプログラムがコンピュータ上で実行される場合に本発明による方法手順を実行する。
【００２５】
以下では本発明の実施例を添付の図面に基づき詳細に説明する。
【図面の簡単な説明】
【００２６】
【図１】本発明の実施形態により得られた報酬と、非特許文献１の方法による対応する報酬とを比較して示す線図である。
【図２】本発明の方法の実施形態により得られた報酬と、品質関数の統計的不確定性を考慮しない方法により得られた報酬とを比較して示す線図である。
【発明を実施するための形態】
【００２７】
以下、本発明を技術システムの例で説明する。この技術システムは状態空間Ｓと活動空間Ａにより特徴付けられる。状態空間とは、技術システムの運転中にこの技術システムを特徴付けるパラメータの形にある多数の離散的または連続的状態の集合である。ガスタービンの場合、これらのパラメータは例えば、供給される燃料量またはタービンのうなりである。活動空間は、技術システムで実行可能な活動を表し、この活動により技術システムの状態を変化することができる。活動は技術システムの調整量の変化であっても良く、例えばガスタービンの案内翼の位置変化等である。
【００２８】
技術システムのダイナミクスは、ここに記載した実施形態ではマルコフ決定プロセスとして遷移確率分布P_T：Ｓ×Ａ×Ｓ→［０，１］により特徴付けられる。この遷移確率分布は、技術システムの目下の状態、目下の状態で実行された活動ならびにそこから生じた技術システムの連続状態に依存する。ここに説明する本発明の方法の実施形態では、トレーニングデータに基づきコンピュータ支援で活動選択ルールが学習される。この活動選択ルールは一般的に、技術システムの所与の状態においてはどの活動を優先的に実行すべきかを指示する。活動選択ルールはここで決定論的であっても良い。すなわち特定の活動がルールによって設定されても良い。しかし活動選択ルールは確率論的であっても良い。すなわち活動選択ルールが、実行すべき活動の確率分布を状態に基づいて指示しても良い。本発明の方法の目的は、期待されるパフォーマンスの点で必ずしも最適ではないが、統計的に活動選択ルールへの最低の要求を満たす、いわば確実性の点で最適化された活動選択ルールを学習することである。このようにして、期待される最大パフォーマンスの最適基準は満たさないが、保証されるパフォーマンスを最大にする活動選択ルールを学習することができる。
【００２９】
ここに説明する本発明の変形実施形態は非特許文献１による方法に基づく。しかし本発明の方法は格段に計算効率がよい。なぜなら、活動選択ルールの不確定性を決定する変数間の相関を考慮しないからである。これについては下でさらに詳細に説明する。
【００３０】
まず従来技術による強化学習方法について説明する。ここでは活動選択ルールが対応する最適基準に基づいて学習される。ここで最適基準は対応する評価Ｒによって表される。この評価は、状態、この状態で実行された活動ａ、および連続状態ｓ’に対するものであり、実行された活動ａが技術システムの最適運転の点でどの程度の価値があるものかを指示する。最適運転は、注目する技術システムに応じて任意に設定することができ、例えばこのような運転に対する基準は、「技術システムの損傷または破壊に繋がるような状態が発生しない」、または「技術システムの運転で理想効率が達成される」である。ガスタービンでは最適運転を例えば、タービンにうなりが発生せずに高い効率が達成されたことに

より特徴付けることができる。

【００３１】

強化学習では活動選択ルールにしたがい、マルコフ決定プロセスＭ （Ｓ，Ａ，$Ｐ_{Ｔ}$，Ｒ）、ただし状態空間Ｓ、活動空間Ａならびに確率分布$Ｐ_{Ｔ}$：Ｓ×Ａ×Ｓ→［０，１］を前提にして、どの活動が技術システムの最適運転に至るかが求められる。ここでは各状態、この状態で実行される活動、およびそこから生じる、報酬関数Ｒ：Ｓ×Ａ×Ｓ→Ｒをともなう連続状態が評価される。ここで最適運転は、いわゆる価値関数の最大値により記述される。これは次式のとおりである。

【００３２】

【数８】

$$V^{\pi}(s)=\sum_{s'}P(s'|s,a)\left[R(s,a,s')+\gamma V^{\pi}(s')\right]$$

【００３３】

この価値関数は、将来の評価の予想される非連続和であり、γ∈［０，１］が非連続因子である。ここでは通例、いわゆるＱ関数$Ｑ^{\pi}$（ｓ，ａ）が使用され、このＱ関数は状態ｓでの活動ａの選択、およびそれに続く活動選択ルールπの実施の後で予想される非連続報酬を表す。ここで最適活動選択ルールに対するＱ関数$Ｑ^{\pi}$＝$Ｑ^{*}$は、いわゆるベルマン最適方程式の解により与えられる。これは次式のとおりである。

【００３４】

【数９】

$$Q^{*}(s,a)=E_{s'}\left[R(s,a,s')+\gamma V^{*}(s')\right]=E_{s'}\left[R(s,a,s')+\gamma\max_{a'}Q^{*}(s',a')\right]$$

【００３５】

ここで$Ｅ_{Ｓ'}$は予想値である。$Ｑ^{*}$に基づき、最適活動選択ルールに対しては$π^{*}$（ｓ）＝argmax_{a} $Ｑ^{*}$（ｓ，ａ）が成り立つ。ここで$π^{*}$は決定論的活動選択ルールである。しかし上にすでに述べたように、活動選択ルールは統計的活動選択ルールπ（ａ｜ｓ）として構築することもでき、これは状態ｓにおいて活動ａを選択するための確率を提供する。

【００３６】

上記のベルマン最適方程式は、従来技術から十分に公知のベルマン反復式より解かれる。これについて下にさらに説明する。以下でＴはベルマン演算子として定義され、各任意の品質関数Ｑに対して次のとおりである。

【００３７】

【数１０】

$$(TQ)(s,a)=E_{s'}\left(R(s,a,s')+\gamma\max\nolimits_{a'}Q(s',a')\right)$$

【００３８】

以下に説明する本発明の実施形態では、統計的不確定性が付加的に注目される。この統計的不確定性は技術システムの測定の不確定性から生じるものであり、技術システムのための活動選択ルールを決定するトレーニングデータとして使用される。

【００３９】

この統計的不確定性は注目するＱ関数、すなわち学習された活動選択ルールの不確定性を引き起こす。強化学習に存在する不確定性は、技術システムの真の特性についての無知から生じるものである。すなわち技術システムの基礎となる、真のマルコフ決定プロセスから生じる。技術システムに関してトレーニングデータの形でより多くの観察が存在すれば、マルコフ決定プロセスに関してより多くの情報が得られる。偶然性が大きければ、所与数の観察に対するマルコフ決定プロセスを基準にしてより多くの不確定性が残る。

【００４０】
トレーニングデータに基づく測定の不確定性、すなわち１つの状態から、活動を適用した次の状態への変遷、およびこれと結び付いた評価は、以下に説明する本発明の変形実施形態ではＱ関数に、不確定性伝播によって伝播する。不確定性伝播の原理は、不確定性のガウス伝播またはガウスエラー伝播とも称され、従来技術から十分に公知であり、それぞれ推定された点を中心にする一次のテイラー展開に基づくものである。非特許文献１に記載された方法によれば、関数ｆ（ｘ）、ただしｆ：Ｒ^{m}→Ｒ^{n}の不確定性は、独立変数ｘの不確定性が所与である場合、以下の共分散に基づいて示される。
Ｃｏｖ（ｆ）＝Ｃｏｖ（ｆ，ｆ）＝ＤＣｏｖ（ｘ）Ｄ^{T}
ここで
【数１１】

$$D_{i,j}=\frac{\partial f_i}{\partial x_j}$$

は、その独立変数ｘによるｆのヤコビ行列である。Ｃｏｖ（ｘ）＝Ｃｏｖ（ｘ，ｘ）により、独立変数ｘの共分散が示され、この共分散はさらにｘの不確定性の関数である。関数ｆは、対称性で正の規定共分散および不確定性Ｃｏｖ（ｆ）を使用する。ここで非特許文献１の方法は、ｍ番目のベルマン反復ステップで、この反復ステップでのＱ関数Ｑ^{m}、遷移確率Ｐ、および評価Ｒに依存する完全な共分散行列が計算されるという欠点を有する。各反復ステップにおいて共分散行列を完全に計算することは面倒であり、非常に大きな計算時間を必要とする。
【００４１】
本発明によれば、共分散マトリクスの非対角要素を無視することにより、すなわちゼロにセットすることにより、非特許文献１の方法を計算的に格段に効率良く構築できることを認識した。これは、共分散マトリクスを決定する変数の不確定性の相関、すなわちＱ関数Ｑ^{m}、遷移確率Ｐおよび評価Ｒ間の相関は無視することができるという仮定に相当する。このように近似しても、なお非常に良好な活動選択ルールを学習することができ、このことは本発明者により実験により証明された。本発明の方法の利点は、その計算時間が非特許文献１の方法の場合よりも何倍も小さいことである。以下に本発明の方法を、実施例に基づいて詳細に説明する。
【００４２】
非特許文献１の方法と同じように、不確定性伝播ないしガウスエラー伝播は、測定の不確定性、すなわち遷移確率と評価の不確定性を、Ｑ関数へ、ひいては活動選択ルールへ伝播させるのに使用することができる。共分散マトリクスが対角要素だけを含むという近似に基づいて、関数値ｆ（ｘ）、ただしｆ：Ｒ^{m}→Ｒ^{n}を分散として記述することができる。
【００４３】
【数１２】

$$(\sigma f)^2=\sum_i\left(\frac{\partial f}{\partial x_i}\right)^2(\sigma x_i)^2$$

【００４４】
このように変数間の相関を無視して不確定性を近似的に考慮することは、ベルマン反復式の次式により表されるｍ番目の反復ステップにおいて
【数１３】

$$Q^m(s,a)=\sum_{s'}P(s'|s,a)\left[R(s,a,s')+\gamma V^{m-1}(s')\right]$$

Ｑ関数における次の不確定性となる。

【００４５】

【数１４】

$$(\sigma Q^{m}(s,a))^{2} = \sum_{s'} (D_{QQ})^{2} (\sigma V^{m-1}(s'))^{2} + \sum_{s'} (D_{QP})^{2} (\sigma P(s'|s,a))^{2} + \sum_{s'} (D_{QR})^{2} (\sigma R(s,a,s'))^{2},$$

$$(D_{QQ}) = \gamma P(s'|s,a),$$

$$(D_{QP}) = R(s,a,s') + \gamma V^{m-1}(s'),$$

$$(D_{QR}) = P(s'|s,a)$$

【００４６】

上記の方程式では、確率論的活動選択ルールπの一般的場合が仮定されている。ここでπ（ａ｜ｓ）は状態ｓにおける活動ａの選択の確率を表す。この表記法は決定論的活動選択ルールπ_{d}を記述するのにも使用することができる。このような場合、π_{d}（ｓ）＝ａであればπ（ａ｜ｓ）　１が成り立ち、π_{d}（ｓ）≠ａであればπ（ａ｜ｓ）　０が成り立つ。所与の活動選択ルールの判定または評価に関し、上記のパラメータV^{m}（ｓ）と（σV^{m}（ｓ））^{2}は確率論的活動選択ルールπに対しては次のとおりである。

【００４７】

【数１５】

$$V^{m}(s) = \sum_{a} \pi(a|s) Q^{m}(s,a),$$

$$(\sigma V^{m}(s))^{2} = \sum_{a} \pi(a|s) (\sigma Q^{m}(s,a))^{2}$$

【００４８】

これに対し、決定論的活動選択ルールに対するパラメータは次のとおりである。

【００４９】

【数１６】

$$V^{m}(s) = Q^{m}(\pi(s),a),$$

$$(\sigma V^{m}(s))^{2} = (\sigma Q^{m}(\pi(s),a))^{2}$$

【００５０】

上記のベルマン最適方程式による活動選択ルールの反復計算の場合、ベルマン反復式のｍ次の反復ステップにおける最適活動選択ルールのＱ関数Ｑ＊に対するＶまたはσＶは次のとおりである。

【００５１】

【数１７】

$$V^{m}(s) = \max Q(s,a)$$

$$(\sigma V^{m}(s))^{2} = (\sigma Q(s, \arg\max_{a} Q(s,a)))^{2}$$

【００５２】

本発明によれば、上記の不確定性伝播がベルマン反復式と平行して使用され、各反復ステップQ^mとσQ^mで更新される。

【００５３】

ここで対応する推定子が、遷移確率Ｐに対し、また不確定性σＰまたはσＲによる評価Ｒに対し使用される。これについては後で詳細に説明する。最初、Ｑ関数Q^0による反復では、対応する不確定性σQ^0により開始される。ここでは例えばQ^0　０、σQ^0　０とすることができる。

【００５４】

上記の反復式が収束する場合、対応する不確定性σQ^*を備えるQ^*の固定値に達する。この情報は、統計的不確定性を考慮する後続のＱ関数を得るために使用することができる。

【００５５】

【数１８】

$$Q_u^*(s,a)=Q^*(s,a)-\xi\sigma Q^*(s,a)$$

【００５６】

この不確定性を考慮するＱ関数は、Ｐ（ζ）の保証確率を備える予想パフォーマンスを提供する。ただし、活動ａが状態ｓで実行され、続いて活動選択ルール$\pi^*(s)=\mathrm{argmax}_a Q^*(s,a)$が遵守されるという条件の下である。ここでは$Q^*_u$、すなわち$\pi_u(s)=\mathrm{argmax}_a Q^*_u(s,a)$に基づく活動選択ルールは、保証されたパフォーマンスを一般的には改善しないことに注意すべきである。なぜならQ^*_uは反復ステップで不確定性だけに注目するからである。一般的にQ^*_uは活動選択ルールπ_uのＱ関数を表さない。このことは不整合につながる。不確定性に関する知識を、保証されたパフォーマンスを最大にするのに利用するため、ベルマン反復式の各反復ステップで活動選択ルールの更新の際に、不確定性を考慮しなければならない。

【００５７】

したがってここに説明する本発明の実施形態では、ｍ次のベルマン反復ステップにおける最適活動選択ルールが$Q^m(s,a)$に基づいて計算されるのではなく、修正されたＱ関数$Q^m(s,a)-\zeta\sigma Q^m(s,a)$に基づいて計算される。すなわちここに説明する本発明の実施形態では、活動選択ルールの計算が次式に基づいて行われる。

【００５８】

【数１９】

$$\forall s: a_{s,\max}=\arg\max_a Q^m(s,a)-\xi\sigma Q^m(s,a)$$

【００５９】

したがって反復式の後続のステップでは、$a_{s,max}$が$\mathrm{argmax}_a Q(s,a)$の代わりに使用され、Q^{m-1}とσQ^{m-1}に対する適値が決定される。

【００６０】

このようにして、パラメータζに依存する信頼値について最適である活動選択ルールが得られる。すなわちその最小パフォーマンスが所与の確率に依存して保証される活動選択ルールが得られる。したがって形式的には活動選択ルールπが、保証パフォーマンスＺ（ｓ，ａ）の最大化により次式のように得られる。ここでは次式が当てはまる。

【００６１】

【数２０】

$$\forall s,a: P\left(\overline{Q}^{\pi}>Z(s,a)\right)>P(\xi)$$

【００６２】

ここでQ^{π}（ただしＱの上にバーあり）はπの真のＱ関数を、Ｐ（ζ）は前もって特定

された固定の確率を表す。したがってパフォーマンスＺはＱ$^{\pi}{}_{u}$により近似され、次式により解かれる。

【００６３】

【数２１】

$$\pi^{\xi}(s)=\arg\max_{\pi}\max_{a}Q_{u}^{\pi}(s,a)=\arg\max_{\pi}\max_{a}Q^{\pi}(s,a)-\xi\sigma Q^{\pi}(s,a)$$

ただし、Ｑ$^{\pi}$がπの有効Ｑ関数であるという条件の下である。

【００６４】

決定論的活動選択ルールの場合はこのようにして、ベルマン反復式の枠内で次の確率最適活動選択ルールが得られる。

【００６５】

【数２２】

$$\pi^{m}(s)=\arg\max_{a}Q^{m}(s,a)-\xi\sigma Q^{m}(s,a)$$

【００６６】

この活動選択ルールは各反復ステップにおいて最適の活動を、特定の状態でのＱ値の最大値を基準にするのではなく、Ｑ値の最大値から重み付けしたその不確定性を減じたものを基準にして形成する。ここで重み付けはパラメータζに基づき適切に設定される。

【００６７】

上記のベルマン反復式に基づく決定論的活動選択ルールが、収束することを保証することはできない。とりわけ、活動選択ルールを発振させる、すなわち対応するＱ関数を収束させない２つの作用が存在する。第１の作用は、すでに非特許文献１に記載されており、ζσＱ（ｓ，ａ）≠π（ｓ）よりも大きなζσＱ（ｓ、π（ｓ））に基づくものであり、πが求める活動選択ルールの場合、ζ＞０が成り立つ。これは、Ｒ（ｓ，π（ｓ）、ｓ’）とＶ（ｓ７）＝Ｑ（ｓ’π（ｓ））が、Ｒ（ｓ，ａ，ｓ’）≠π（ｓ）とＶ（ｓ’）より強く相関しているためである。というのも、状態ｓが比較的後で発生するたびに活動π（ｓ）の選択を価値関数が暗示するからである。特定の状態ｓで活動選択ルールをπからπ’に、Ｑ（ｓ、π（ｓ））－ζσＱ（ｓ、π（ｓ））＜Ｑ（ｓ、π’（ｓ））－ζσＱ（ｓ，π’（ｓ））という条件のため切り換えると、Ｑ（ｓ、π’（ｓ））の不確定性が比較的大きくなり、したがって次の反復ステップで再び始めに戻ることがある。

【００６８】

すでに述べたように、Ｑ値と、発生する活動の対応する不確定性とに特定の状況が存在する場合、発振を引き起こす第２の作用が存在する。そのような状況の例は、２つの活動ａ$_{1}$とａ$_{2}$が１つの状態ｓで類似のＱ値を有するが、不確定性は異なる場合である。これは、ａ$_{1}$が比較的大きな不確定性を有するが、真のマルコフ決定プロセスに対してはより良好な場合である。不確定性を考慮する活動選択ルールを更新するステップでは、π^{m}により、不確定性が最小である活動ａ$_{2}$が選択されるようになる。しかし場合によっては、この活動が比較的劣であるとランク付けられている事実が、変更された活動選択ルールπ^{m}（活動ａ$_{2}$を選択する活動選択ルール）に対する価値関数が更新される場合に次の反復ステップで際立つことがある。したがって活動選択ルールの更新の際に次のステップで、状態ｓで活動ａ$_{1}$が選択されるように活動選択ルールが変更される。なぜならＱ関数は、活動ａ$_{2}$が活動ａ$_{1}$より劣っていることを反映しているからである。Ｑ関数の次の更新後に、両方の活動に対する値は類似するようになる。なぜなら価値関数がａ$_{1}$の選択を暗示し、ａ$_{2}$の劣った作用が関数Ｑ（ｓ，ａ$_{2}$）を一度調整するからである。したがって活動ａ$_{1}$とａ$_{2}$との間で発振が生じる。ここでは、非特許文献１に記載の方法では上記２つの作用が発生するが、本発明の方法では第２の作用だけが関連することに注意すべきである。これは、Ｑ関数と報酬との間の共分散が考慮されていないためである。

【００６９】

上記の非共分散の問題を解決するために、とくに好ましい実施形態では、確率論的活動選択ルールが適切な更新ステップにより上記の活動$a_{s,max}$に基づいて決定される。$\zeta > 0$に対しては確率最適活動選択ルールが確率論的なものであることは直観的に自明である。なぜなら将来の報酬が低下するリスクを最小にすることが試行されるからである。

【００７０】

ここに説明した本発明の変形実施形態では、確率の同じ活動により初期設定される確率論的活動選択ルールが使用される。各反復ステップで、$Q^{\pi}{}_{u}$にしたがい最適の活動の確率が高められる。一方、他のすべての活動の確率は次式に基づき低下される。

【００７１】

【数２３】

$$\forall s,a:\pi^{m}(a|s)=\begin{cases}\min(\pi^{m-1}(a|s)+1/t,1), & falls\ \ a=a_{Q_u}(s)\\ \dfrac{\max(1-\pi^{m-1}(s,a_{Q_u}(s))-1/t,0)}{1-\pi^{m-1}(s,a_{Q_u}(s))}\pi^{m-1}(a|s), & sonst\end{cases}$$

【００７２】

ここで

$a_{Q_u}(s)$

はQ_uによる最適の活動を表す。すなわち

【数２４】

$$a_{Q_u}(s)=\arg\max_a Q(s,a)-\xi\sigma Q(s,a)$$

が当てはまる。

【００７３】

調和的に減少する変化率に基づき、可能なすべての活動選択ルールの収束と到達可能性が保証される。ここでは収束が保証される他に、本発明により実施された実験で、確率論的活動選択ルールが決定論的活動選択ルールよりも良好な結果を提供することが示された。

【００７４】

従来のベルマン反復式の時間複雑性はO（｜S｜^{2}｜A｜）にある。本発明の方法では、Q関数の不確定性σQを更新するステップが挿入され、このステップも同様にO（｜S｜^{2}｜A｜）の時間複雑性を有する。したがってこの方法全体が、O（｜S｜^{2}｜A｜）の時間複雑性を有する。非特許文献１による方法は完全な共分散マトリクスを計算し、時間複雑性をO（（｜S||A｜）^{2}ｌｏｇ（｜S||A｜）とO（（｜S||A｜）^{2376}）の間に、共分散マトリクスの更新時に挿入する。そのため従来のベルマン反復式よりも時間複雑性が大きくなる。標準ベルマン反復式のメモリスペース複雑性は遷移確率Pとステイされた評価Rにより決定され、これらはそれぞれO（（｜S｜^{2}｜A｜）のメモリスペースを必要とする。Q関数はO（｜S||A｜）のメモリスペースを必要とする。したがって標準ベルマン反復式の全メモリスペースはO（｜S｜^{2}｜A｜）である。不確定性をインプリメンテーションすることにより、σPとσRに対するO（｜S｜^{2}｜A｜）の複雑性と、σQに対するO（｜S||A｜）の複雑性が挿入される。したがって全体のメモリスペース複雑性はO（｜S｜^{2}｜A｜）において同じである。これとは異なり、非特許文献

１による方法は、完全な共分散マトリクスのためのメモリスペースを必要とする。この完全な共分散マトリクスは、部分行列Ｃｏｖ（Ｑ），Ｃｏｖ（Ｑ，Ｐ），Ｃｏｖ（Ｑ，Ｒ），Ｃｏｖ（Ｐ），Ｃｏｖ（Ｐ，Ｒ）およびＣｏｖ（Ｒ）からなる。そのためメモリスペース複雑性はＯ（｜Ｓ｜^{5}｜Ａ｜^{3}）となる。したがって時間複雑性もメモリスペース複雑性も、ここに説明した方法では、非特許文献１の方法の場合よりも格段に小さくなることが明白である。

【００７５】

すでに上に示したように、確率最適活動選択ルールを求めるための計算は、トレーニングデータによる遷移確率Ｐと評価Ｒの推定に基づく。ここでは例えば、ＰとＲに対する一般的な推定を、発生する状態の相対的頻度を用い、トレーニングデータに基づき使用することができる。この場合、遷移確率は多項分布としてモデル化され、これに基づき不確定性が次のように計算される。

【００７６】

【数２５】

$$\left(\sigma P\left(s'\middle|s,a\right)\right)^2=\frac{P\left(s'\middle|s,a\right)\left(1-P\left(s'\middle|s,a\right)\right)}{n_{sa}-1}$$

【００７７】

ここでＰ（ｓ’｜ｓ，ａ）は、状態ｓと活動ａを前提とする連続状態ｓ’の相対頻度に相当する。さらにn_{sa}は、状態活動ペア（ｓ，ａ）に基づく連続状態への、観察された遷移の数を表す。これらの情報はすべてトレーニングデータから由来する。

【００７８】

同じようにして評価を、正規分布を前提にしてモデル化することができる。この場合、遷移（ｓ，ａ，ｓ’）で観察されたすべての評価の平均値が評価に対する予想値として使用される。したがい評価に対する不確定性は次のようになる。

【００７９】

【数２６】

$$\left(\sigma R\left(s,a,s'\right)\right)^2=\frac{\mathrm{var}\left(R\left(s,a,s'\right)\right)}{n_{sas'}-1}$$

【００８０】

ここで分散の分子の表現は、トレーニングデータに基づいてモデル化された正規分布に相当する。さらに$n_{sas'}$は、観察された遷移（ｓ，ａ，ｓ’）の数である。

【００８１】

相対的頻度に基づく上記の遷移確率の推定は、通例、良好な結果を生む。しかし対応する不確定性推定は、トレーニングデータより少数の観察しか存在しない場合に問題である。例えば特別の遷移が２度、２回の試行で観察されれば（すなわち（$n_{sas'}=n_{sa}=2$)が成り立てば)、それ自身の不確定性はσＰ(ｓ’｜ｓ，ａ）＝０となる。これにより、観察が少数の場合には、不確定性がしばしば過度に低くランク付けられる。

【００８２】

遷移確率の決定のためによく利用される数式の代わりに、ベイズの推定を使用することもできる。ここでは状態s_iおよび連続状態s_kに対するパラメータ空間Ｐ（s_k｜s_i，a_j）についてのアプリオリ分布として、以下の密度を備えるディリクレ分布が使用される。

【００８３】

【数２７】

$$\Pr\left(P\left(s_1,\left|s_i,a_j\right.\right),\ldots,P\left(s_{|S|}\left|s_i,a_j\right.\right)\right)_{\alpha_{ij1},\ldots,\alpha_{ij|S|}}=\frac{\Gamma\left(\alpha_{ij}\right)}{\prod_{k=1}^{|S|}\Gamma\left(\alpha_{ijk}\right)}\prod_{k=1}^{|S|}P\left(s_k\left|s_i,a_j\right.\right)^{\alpha_{ijk}-1}$$

ここでは

【数２８】

$$\alpha_{ij}=\sum_{k=1}^{|S|}\alpha_{ijk}$$

が成り立つ。ディリクレ分布は、次のアポステリオリパラメータを備える、いわゆる「共役プリオー（conjugate prior）」である。

【００８４】

【数２９】

$$\alpha_{ijk}^{d}=\alpha_{ijk}+n_{s_i a_j s_k},\alpha_{ij}^{d}=\sum_{k=1}^{|S|}\alpha_{ijk}^{d}$$

ここで

$$n_{s_i a_j s_k}$$

は、トレーニングデータにしたがい活動ａ$_{j}$を実施した際の、ｓ$_{i}$からｓ$_{k}$への遷移の数である。アポステリオリ分布の予想値を推定子として使用することにより、すなわち確率をＰ（ｓ$_{k}$｜ｓ$_{i}$，ａ$_{j}$）＝α$^{d}_{ijk}$／α$^{d}_{ij}$と推定することにより、Ｐに対する不確定性は次のようになる。

【００８５】

【数３０】

$$\sigma P\left(s_k\left|s_i,s_j\right.\right)=\frac{\alpha_{ijk}^{d}\left(\alpha_{ij}^{d}-\alpha_{ijk}^{d}\right)}{\left(\alpha_{ij}^{d}\right)^2\left(\alpha_{ij}^{d}+1\right)}$$

α$_{ijk}$はディリクレ分布のパラメータである。α$_{ijk}$＝０と選択することにより、遷移確率の上記一般的なモデル化と比較して、同等の推定とわずかに小さな不確定性が得られる。他方、α$_{ijk}$＝１と選択することにより、１つの状態から別のすべての状態へのすべての遷移が同じ確率である分布が生じる。

【００８６】

α$_{ijk}$＝０とα$_{ijk}$＝１の選択はそれぞれ、ほとんどの適用に適しない極値である。したがって本発明のとくに好ましい変形実施形態では、ディリクレ分布のパラメータが次のように設定される。

【００８７】

【数３１】

$$\alpha_{ijk}=\frac{m}{|S|}$$

【００８８】

ここでｍはすべての状態活動ペアの予想される連続状態の平均数であり、｜Ｓ｜は状態の総数である。α$_{ijk}$を好ましく選択することにより、状態パラメータｍの状態空間の部分集合にわたり、最大のエントロピーによりアプリオリ確率を近似することができる。このようにして確率の大部分が、実際に観察された状態ｍの部分集合に分散され、他の（観察されない）すべての連続状態の確率が非常に小さくなる。α$_{ijk}$＝１によるアプリオリ分布と比較して、観察された連続状態に対して、観察されなかった連続状態に対するものより高い確率を達成するために、実際に観察された連続状態に対して少数の観察しか必要ない。同時に不確定性の推定が、一般的に使用される数式の場合より極端でない。な

ぜなら同じ観察が２度行われても、不確定性がゼロにはならないからである。

【００８９】

本発明の方法の実施形態が、いわゆる「ウェットチキン（Wet Chicken）」ベンチマーク問題でテストされた。オリジナルのウェットチキン問題では、長さｌ、流速ｖ＝１の一次元の流れに沿ってパドルするカヌー漕手が考察される。流れの位置ｘ＝ｌには滝がある。位置ｘ＝０から出発してカヌー漕手は、できるだけ滝に接近するが、滝からは落下しないことを試みる。カヌー漕手が滝から落下すると、彼は再び位置ｘ＝０から開始しなければならない。報酬または評価は、滝に接近するとともに線形に上昇し、ｒ＝ｘにより表される。カヌー漕手は、流される、自分の位置を保持する、または流れに逆らって漕ぐなどの手段を有する。流れの渦がパラメータｓ＝２．５であると、状態の確率論的遷移が生じる。カヌー漕手が自分の現在位置（河の流れも考慮して）で活動を実施した後、彼の新たな位置はｘ’＝ｘ＋ｎにより与えられる。ここでｎ∈［－ｓ，ｓ］が同じように分散されたランダム値である。ここで考察する２次元のウェットチキン問題が、幅ｗだけ拡張あれる。したがって、カヌー漕手に対しては付加的に２つの活動が可能である。カヌー漕手はカヌーを一方では右に、他方では左に１単位だけ移動することができる。カヌー漕手の位置は（ｘ、ｙ）として示され、スタート位置は（０，０）である。流速ｖと渦の量ｓはｙに依存し、ｖ＝３ｙ／ｗと、ｓ＝３．５－ｖが成り立つ。実験では、離散的問題が考察された。すなわちｘとｙの値は常に次の整数値に丸められた。河の流速は左岸でゼロであるが、そこで渦の量が最大である。一方、河の右岸には渦がないが、逆に漕ぐための流速は最高である。

【００９０】

上記の問題に基づいて、対応する活動選択ルールが本発明の方法により学習された。ここで図１と２に示された実験１００で、河において可能な状態が考察された。すなわちカヌー漕手は河の中で１０×１０の可能な位置を取ることができる。別の実験では５×５または２０×２０の状態が考察された。固定数の観察が、状態空間のランダムな調査により発生された。発生された観察が本発明による確率最適の活動選択ルールの決定のための入力量（すなわちトレーニングデータ）として使用された。ここでは非連続因子ζが０．９５に設定された。活動選択ルールの決定後、このルールが１００のエピソードに関して、それぞれ１０００ステップによりテストされた。

【００９１】

図１は、テストされた活動選択ルールの結果であり、１００の試行にわたり平均されている。ここで図１には、活動選択ルールの平均報酬が、活動選択ルールを学習するための使用された観察数の関数として示されている。観察は横軸にＯにより示されており、平均報酬は縦軸にＡＲとして示されている。直線Ｌ１は、遷移確率を推定するための一般的数式に対する安全パラメータζ＝０．５についての結果を示し、直線Ｌ２は、ベイズの数式に対するζ＝０．５についての結果を示し、直線Ｌ３は一般的数式に対するζ＝１についての結果を示し、直線Ｌ４はベイズの数式に対するζ＝１についての結果を示す。比較のため不確定性を考慮しない（すなわちζ＝０）活動選択ルールの学習結果が、直線Ｌ５により示されている。さらに非特許文献１による完全な共分散マトリクスに基づく活動選択ルールの学習が示されている。ここで直線Ｌ６は、ζ＝１に対して非特許文献１の方法により学習された活動選択ルールを示し、直線Ｌ７は、ζ＝０．５に対する非特許文献１の方法による結果を示す。簡単にするため、確率論的活動選択ルールだけが、ζ＝０を除いて考察された。とりわけ図１から、非特許文献１による方法のパフォーマンスは確かに高いが、本発明の方法についても良好な結果が達成されていることが分かる。このことは平均報酬の高い場合において、とくに観察数が多いときに反映されている。さらに統計的不確定性を考慮する方法は、観察数が多い場合に、活動選択ルールにおいて不確定性を考慮しない方法よりも良好である。

【００９２】

図２は、活動選択ルールの頻度を、学習された１０００の活動選択ルールに対する平均報酬の関数として示すヒストグラムである。横軸に沿って平均報酬ＡＲが、縦軸に沿って

対応する平均報酬により学習された活動選択ルールの数ＮＰがプロットされている。このヒストグラムで、実線Ｌ８はζ＝０により（すなわち不確定性を考慮せずに）学習された活動選択ルールを、破線Ｌ９は本発明の方法にしたがいζ＝１により学習された活動選択ルールを、点線Ｌ１０は本発明の方法に従いζ＝２により学習された活動選択ルールを示す。各活動選択ルールを形成するために、４×１０^{4}の観察が使用された。図２から、本発明の方法により学習された活動選択ルールは、平均報酬の大きい領域で顕著な最大頻度を有することが分かる。この最大頻度は、不確定性を考慮しない活動選択ルールでは小さい。したがって本発明の方法により発生された、平均報酬の小さい活動選択ルールは、不確定性を考慮しない対応する活動選択ルールと比較して小さい頻度を有する。したがい不確定性の考慮により、報酬の小さい活動選択ルールの量が低減され、期待されるパフォーマンスが上昇する。

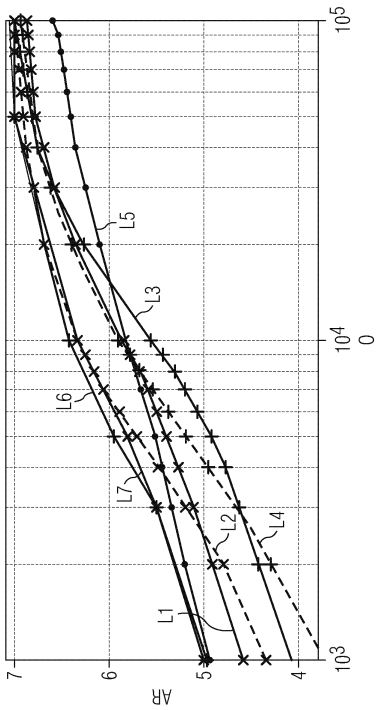
【００９３】

非特許文献１の方法に対する本発明の方法の格別の利点は、計算時間が格段に短いことである。５×５の状態を備えるウェットチキン問題に対し、非特許文献１に記載の方法では選択ルールを形成するための計算時間が５．６１ｓであった。これに対して本発明の方法は０．０００２ｓしか必要としなかった。１０×１０の状態を備えるウェットチキン問題では、非特許文献１の方法の計算時間は$１．１\times１０^{3}$ｓであった。これに対して本発明の方法は０．０３４ｓしか必要としなかった。２０×２０の状態を備えるウェットチキン問題に対しては、そこから生じる計算時間とメモリスペースが非常に大きいため、非特許文献１に記載の方法により活動選択ルールを求めることができなかった。これに対して本発明の方法は活動選択ルールを発生するのに１．６１ｓしか必要としなかった。

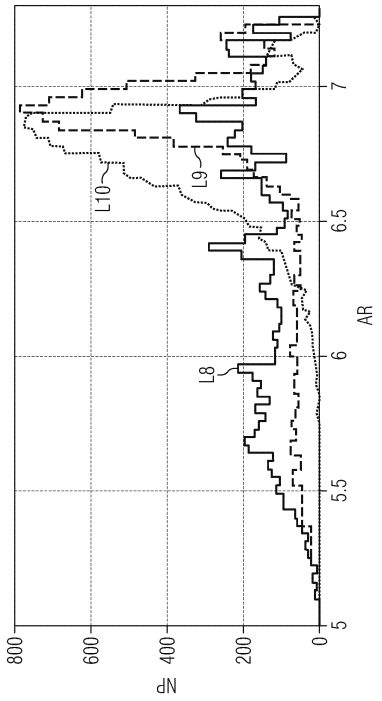
【００９４】

上記のウェットチキン実験は、公知の方法に対する本発明の方法の優れたパフォーマンスを明確にするためにだけ用いるものである。本発明に基づき、技術システムを制御または調整するための方法が使用される。本発明の方法をテストするために、ガスタービン制御のシミュレーションも実行された。このシミュレーションについても本発明の方法は、計算時間の短い良好なパフォーマンスを示した。

【図１】


AR
7
6
5
4
10^5
10^4
10^3
0
L1
L2
L3
L4
L5
L6
L7

【図２】


NP
800
600
400
200
0
AR
5
5.5
6
6.5
7
L8
L9
L10

フロントページの続き

(74)代理人 100112793
弁理士 高橋 佳大
(74)代理人 100128679
弁理士 星 公弘
(74)代理人 100135633
弁理士 二宮 浩康
(74)代理人 100156812
弁理士 篠 良一
(74)代理人 100114890
弁理士 アインゼル・フェリックス＝ラインハルト
(72)発明者 アレクサンダー ハンス
ドイツ連邦共和国 ミュンヘン ヴィンフリードシュトラーセ 5ツェー
(72)発明者 シュテフェン ウードルフト
ドイツ連邦共和国 アイヒェナウ ハービヒトシュトラーセ 2

審査官 多胡 滋

(56)参考文献 特表平１０－５０４６６７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ９９／００
Ｇ０６Ｎ ７／００
Ｇ０５Ｂ １３／０２
Ｇ０６Ｑ ５０／０４