(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2020年6月18日(18.06.2020)

WIPO | PCT

(10) 国際公開番号
WO 2020/121494 A1

(51) 国際特許分類:
G06N 20/00 (2019.01)

(21) 国際出願番号： PCT/JP2018/045947

(22) 国際出願日： 2018年12月13日(13.12.2018)

(25) 国際出願の言語： 日本語

(26) 国際公開の言語： 日本語

(71) 出願人: 日本電気株式会社 (**NEC CORPORATION**) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

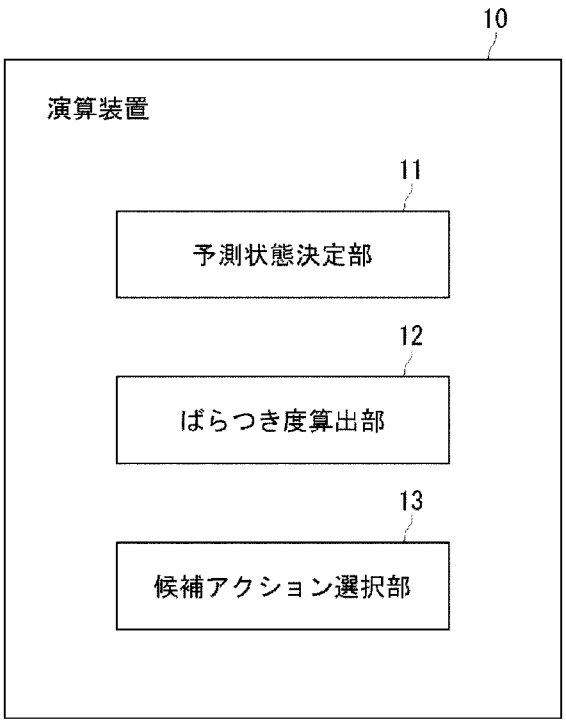
(72) 発明者: 森 達哉(**MORI Tatsuya**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). 平岡 拓也(**HIRAOKA Takuya**); 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP). タンカラットブット(**TANGKARATT Voot**); 〒3510198 埼玉県和光市広沢２番１号 国立研究開発法人理化学研究所内 Saitama (JP).

(74) 代理人: 家入 健(**IEIRI Takeshi**); 〒2210835 神奈川県横浜市神奈川区鶴屋町三丁目３３番８ アサヒビルヂング５階 響国際特許事務所 Kanagawa (JP).

(81) 指定国(表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH,

(54) **Title:** ARITHMETIC DEVICE, ACTION DETERMINATION METHOD, AND NON-TRANSITORY COMPUTER-READABLE MEDIUM STORING CONTROL PROGRAM

(54) 発明の名称： 演算装置、アクション決定方法、及び制御プログラムを格納する非一時的なコンピュータ可読媒体



10 Arithmetic device
11 Predictive state determination unit
12 Dispersion degree calculation unit
13 Candidate action selection unit

(57) **Abstract:** In an arithmetic device (10), a predictive state determination unit (11) determines, using a plurality of transition information units, a plurality of predictive states pertaining to each of a plurality of candidate actions possible in a first state. A dispersion degree calculation unit (12) calculates the degree of dispersion of the plurality of predictive states

WO 2020/121494 A1

[続葉有]

KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国(表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

一 国際調査報告(条約第21条(3))

determined pertaining to each candidate action by the predictive state determination unit (11). A candidate action selection unit (13) selects some of the plurality of candidate actions on the basis of multiple degrees of dispersion calculated by the dispersion degree calculation unit (12).

(57) 要約:演算装置(１０)にて予測状態決定部(１１)は、複数の遷移情報ユニットを用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の予測状態を決定する。ばらつき度算出部(１２)は、予測状態決定部(１１)にて各候補アクションに関して決定された複数の予測状態のばらつき度を算出する。候補アクション選択部(１３)は、ばらつき度算出部(１２)にて算出された複数のばらつき度に基づいて、上記の複数の候補アクションのうちの一部の候補アクションを選択する。

明　細　書

発明の名称 ：

演算装置、アクション決定方法、及び制御プログラムを格納する非一時的なコンピュータ可読媒体

技術分野

[0001] 本開示は、演算装置、アクション決定方法、及び制御プログラムに関する。

背景技術

[0002] 「強化学習」に関して種々の研究が行われている（例えば、非特許文献１）。強化学習の目的の１つは、実環境に対して時系列的に複数の行動（アクション）を行った結果、実環境から得られる「累積報酬」を最大化する、方策（Ｐｏｌｉｃｙ）を学習することである。

先行技術文献

非特許文献

[0003] 非特許文献1：Ｒｉｃｈａｒｄ Ｓ．Ｓｕｔｔｏｎ ａｎｄ Ａｎｄｒｅｗ Ｇ．Ｂａｒｔｏ，“Ｒｅｉｎｆｏｒｃｅｍｅｎｔ Ｌｅａｒｎｉｎｇ：Ａｎ Ｉｎｔｒｏｄｕｃｔｉｏｎ”，Ｓｅｃｏｎｄ Ｅｄｉｔｉｏｎ，ＭＩＴ Ｐｒｅｓｓ，２０１８

発明の概要

発明が解決しようとする課題

[0004] ところで、好適な方策を効率良く学習するためには、実環境の状態についての「状態空間」を効率的に探索する必要がある。

[0005] しかしながら、非特許文献１では探索の重要性について言及されているが、効率的な探索を実現する具体的な技術については開示されていない。

[0006] 本開示の目的は、効率的な探索を実現できる、演算装置、アクション決定方法、及び制御プログラムを提供することにある。

課題を解決するための手段

[0007] 第１の態様にかかる演算装置は、第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定する決定手段と、各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出する算出手段と、前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する選択手段と、を具備する。

[0008] 第２の態様にかかるアクション決定方法は、情報処理装置によって、第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定し、各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出し、前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する。

[0009] 第３の態様にかかる制御プログラムは、第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定し、各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出し、前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する処理を、演算装置に実行させる。

発明の効果

[0010] 本開示により、効率的な探索を実現できる、演算装置、アクション決定方法、及び制御プログラムを提供することができる。

図面の簡単な説明

[0011] [図1]第１実施形態の演算装置の一例を示すブロック図である。

[図2]第２実施形態の演算装置を含む制御装置の一例を示すブロック図である

。

[図3]第２実施形態の演算装置の処理動作の一例を示すフローチャートである。

[図4]第３実施形態の演算装置を含む制御装置の一例を示すブロック図である。

[図5]第３実施形態の演算装置の処理動作の一例を示すフローチャートである。

[図6]演算装置のハードウェア構成例を示す図である。

発明を実施するための形態

[0012] 以下、図面を参照しつつ、実施形態について説明する。なお、実施形態において、同一又は同等の要素には、同一の符号を付し、重複する説明は省略される。

[0013] ＜第１実施形態＞

図１は、第１実施形態の演算装置の一例を示すブロック図である。図１において演算装置（アクション決定装置）１０は、予測状態決定部１１と、ばらつき度算出部１２と、候補アクション選択部１３とを有している。

[0014] 説明の便宜上、あるタイミング（以降、「第１タイミング」と表す）における制御対象の状態を「第１状態」と表す。あるタイミング以降の１つタイミング（以降、「第２タイミング」と表す）における制御対象の状態を「第２状態」と表す。制御対象の状態は、第１状態に応じたアクションが実施された後に第２状態に変化するとする。また、第１状態と、第２状態とは、必ずしも、相互に異なる状態である必要はなく、同じ状態を表していてもよい。以降の説明においては、説明の便宜上、第１状態と、第２状態との異同によらずに、「第１状態から第２状態に変化する」と表すこととする。また、第１タイミング、及び、第２タイミングは、特定のタイミングを表しているわけではなく、相互に異なる２つのタイミングを表している。

[0015] 予測状態決定部１１は、複数の状態遷移情報（遷移情報ユニット）を用いて、第１状態にて可能な複数の「候補アクション」のそれぞれに関して複数

の「予測状態」を決定する。各遷移情報ユニットは、第１状態と、該第１状態におけるアクションとから、該第１タイミング以降（たとえば、第２タイミング）の予測状態を算出するために用いられる。すなわち、各遷移情報ユニットは、各遷移情報ユニットの第１状態を保持しており、該第１状態及びアクションの組み合わせに応じた予測状態を決定する機能を有している。ここで、例えば、各遷移情報ユニットは、あるタイミングでの実環境の状態（実環境状態）と、該あるタイミングで実環境に対して実際に行われたアクションとが関連付けされたセットを含む「履歴情報」に基づいて作成（訓練）される。該セットは、２つの状態と、当該２つの状態間におけるアクションとが関連付けされた情報を表す。

[0016] ばらつき度算出部１２は、予測状態決定部１１にて各候補アクションに関して決定された複数の予測状態の「ばらつき度」を算出する。ここでは、第１状態にて可能な候補アクションは複数存在しているので、複数の候補アクションにそれぞれ対応する複数のばらつき度が算出されることになる。「ばらつき度」は、例えば、分散値である。

[0017] 候補アクション選択部１３は、ばらつき度算出部１２にて算出された複数のばらつき度に基づいて、上記の複数の候補アクションのうちの一部の候補アクションを選択する。例えば、候補アクション選択部１３は、上記の複数の候補アクションのうちで、ばらつき度算出部１２にて算出された複数のばらつき度のうちの最大値に対応する候補アクションを選択する。

[0018] 以上のように第１実施形態によれば、演算装置１０にて予測状態決定部１１は、複数の遷移情報ユニットを用いて、第１状態にて可能な複数の「候補アクション」のそれぞれに関して複数の「予測状態」を決定する。ばらつき度算出部１２は、予測状態決定部１１にて各候補アクションに関して決定された複数の予測状態の「ばらつき度」を算出する。候補アクション選択部１３は、ばらつき度算出部１２にて算出された複数のばらつき度に基づいて、上記の複数の候補アクションのうちの一部の候補アクションを選択する。

[0019] この演算装置１０の構成により、効率的な探索を行うことができる。すな

わち、候補アクションによる第１状態から第２状態への状態遷移が遷移情報ユニットにおいて「訓練の不十分な状態遷移」である場合、その状態遷移の予測状態についての「ばらつき度」が高くなる傾向にある。すなわち、「ばらつき度」は、遷移情報ユニットにおける状態遷移の訓練進捗度を示す指標として用いることができる。また、上記「訓練の不十分な状態遷移」は、上記「履歴情報」に十分に蓄積されていない状態遷移、つまり、実環境において探索が十分でない状態遷移を表していることもある。このため、ばらつき度に基づき候補アクションを選択することによって、探索が十分でない状態遷移（つまり、状態及びアクションの組み合わせ）を積極的に探索することができる。よって、効率的に探索を行なうことができる。さらに、探索が十分でない状態遷移を積極的に探索することができるので、遷移情報ユニットの訓練を効率的に行うことができる。

[0020] ＜第２実施形態＞

第２実施形態は、より具体的な実施形態に関する。

[0021] ＜制御装置の概要＞

図２は、第２実施形態の演算装置３０を含む制御装置２０の一例を示すブロック図である。図２には、制御装置２０の他に、指令実行装置５０及び制御対象６０が図示されている。

[0022] たとえば、制御対象６０が車両である場合に、制御装置２０は、たとえば、エンジンの回転数や、車両の速度や、周囲の状況等の観測値（特徴量）に基づき、ハンドルを右に回す、アクセルを踏む、ブレーキを踏む等のアクションを決定する。指令実行装置５０は、演算装置３０によって決定されたアクションに従いアクセル、ハンドル、または、ブレーキを制御する。

[0023] たとえば、制御対象６０が発電機である場合に、制御装置２０は、たとえば、タービンの回転数や、燃焼炉の温度や、燃焼炉の圧力等の観測値に基づき、燃料の量を増やす、燃料の量を減らす等のアクションを決定する。指令実行装置５０は、制御装置２０によって決定されたアクションに従い燃料の量を調整するバルブを閉める、あるいは、バルブを開く等の制御を実行する

。

[0024] 制御対象６０は、上述した例に限定されず、たとえば、生産工場や、化学工場であってもよいし、車両の動作や、発電機における動作などをシミュレーションしているシミュレータ等であってもよい。

[0025] 観測値に基づきアクションを決定する処理については、図３を参照しながら後述する。

[0026] 制御装置２０は、後述するような、「処理フェーズ１」、「処理フェーズ２」、及び「処理フェーズ３」を実行する。制御装置２０は、これらの処理を実行することによって、制御対象６０の状態（ｓｔａｔｅ）が、より早期に所望状態に近付くようアクションを決定する。この際に、制御装置２０は、制御対象６０の状態に対して実行するアクションを、方策（Ｐｏｌｉｃｙ）情報と、報酬（ｒｅｗａｒｄ）情報とに基づき決定する。

[0027] 方策情報は、制御対象６０がある状態である場合に、実行可能なアクションを表す。方策情報は、たとえば、該ある状態と、該アクションとが関連付けされた情報を用いて実現することができる。方策情報は、たとえば、該ある状態を与えた場合に該アクションを算出する処理であってもよい。当該処理は、たとえば、ある関数、または、統計的な手法によって算出された、当該ある状態と、当該アクションとの関係性を表すモデルであってもよい。すなわち、方策情報は、上述した例に限定されない。

[0028] 報酬情報は、ある状態が望ましい程度（以降、「報酬程度」と表す）を表す。報酬情報は、たとえば、該ある状態と、該程度とが関連付けされた情報を用いて実現することができる。報酬情報は、たとえば、該ある状態を与えた場合に該報酬程度を算出する処理であってもよい。当該処理は、たとえば、ある関数、または、統計的な手法によって算出された、当該ある状態と、当該報酬程度との関係性を表すモデルであってもよい。すなわち、報酬情報は、上述した例に限定されない。

[0029] 以降の説明においては、説明の便宜上、制御対象６０は、車両や、発電機等（以降、「実環境」と表す）であるとする。あるタイミング（以降、「第

１タイミング」と表す）における制御対象６０の状態を「第１状態」と表す。あるタイミングの次のタイミング（以降、「第２タイミング」と表す）における制御対象６０の状態を「第２状態」と表す。制御対象６０の状態は、第１状態に応じたアクションが実施された後に第２状態に変化するとする。また、第１状態と、第２状態とは、必ずしも、相互に異なる状態である必要はなく、同じ状態を表していてもよい。以降の説明においては、説明の便宜上、第１状態と、第２状態との異同によらずに、「第１状態から第２状態に変化する」と表すこととする。

[0030] 制御装置２０は、複数のタイミングに関して、制御対象６０の観測値を参照しながら、処理フェーズ１乃至処理フェーズ３にて後述するような処理を実行することによって、タイミングごとにアクションを決定する。すなわち、制御装置２０は、第１タイミングに関して処理を実行した後に、第２タイミングに関して処理を実行し、さらに、第２タイミングより後のタイミングに関しても処理を実行する。したがって、第１タイミング、及び、第２タイミングは、特定のタイミングを表しているわけではなく、制御装置２０における処理に関して連続している２つのタイミングを表している。

[0031] （処理フェーズ１）

制御装置２０は、状態遷移情報（後述する）に基づき第１状態である制御対象６０に関して、アクションを実行した後の制御対象６０の第２状態を推定する。制御装置２０は、複数の候補アクションに関して、それぞれ、第２状態を推定する処理を実行する。その後、制御装置２０は、報酬情報を用いて、推定した各第２状態に関する報酬程度を算出する。制御装置２０は、複数の候補アクションの中から、算出した報酬程度が上位の候補アクションのうちの１つアクションを選択する。制御装置２０は、複数の候補アクションの中から、算出した報酬程度が最も大きなアクションを１つ選択してもよい。制御装置２０は、選択したアクションを示す制御指令を、指令実行装置５０へ出力する。

[0032] 上位は、たとえば、報酬程度が最も大きいものから、報酬程度が大きい順

に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。

[0033] ここで、状態遷移情報について説明する。状態遷移情報は、第１状態と、第２状態との間の関係性を表す情報である。状態遷移情報は、第１状態と、第２状態とが関連付けされた情報であってもよいし、第１状態と、第２状態とが関連付けされた訓練データを用いたニューラルネットワーク等の統計的な手法によって算出された情報であってもよい。状態遷移情報は、さらに、第１状態にて実行可能なアクションを表す情報を含んでいてもよく、上述した例に限定されない。

[0034] 指令実行装置５０は、制御装置２０によって制御指令を受け取り、受け取った該制御指令が示すアクションを、制御対象６０に関して実行する。この結果、制御対象６０の状態は、第１状態から第２状態に変化する。

[0035] 説明の便宜上、制御対象６０には、制御対象６０を観測しているセンサー（図示せず）が取り付けられているとする。センサーは、制御対象６０に関して観測した観測値を表すセンサー情報を作成し、作成したセンサー情報を出力するとする。制御対象６０を観測しているセンサーは、複数であってもよい。

[0036] 制御装置２０は、第１状態に関するアクションが実行された後に、センサーによって作成された該センサー情報を受け取り、受け取った該センサー情報に関する第２状態を決定する。制御装置２０は、該第１状態と、該アクションと、該第２状態とが関連付けされた情報（以降、「履歴情報」と表す）を作成する。制御装置２０は、作成した履歴情報を、後述する履歴情報記憶部４１に格納してもよい。

[0037] 処理フェーズ１に関して上述したような処理が、複数のタイミングに関して実行されることにより、後述する履歴情報記憶部４１には、複数のタイミングにおける履歴情報が蓄積される。

[0038] （処理フェーズ２）

制御装置２０は、処理フェーズ１にて蓄積された履歴情報を用いて、状態

遷移情報を更新する（または、作成する）。状態遷移情報を、ニューラルネットワークを用いて作成する場合に、制御装置２０は、上述したような履歴情報に含まれているデータを訓練データとして用いて、当該状態遷移情報を作成する。後述するように、制御装置２０は、たとえば、構成が相互に異なっているニューラルネットワークを用いて、複数の状態遷移情報を作成する。

[0039] （処理フェーズ３）

制御装置２０は、複数の候補アクションについて、候補アクションをそれぞれ対象に関して施した後における第２状態を、状態遷移情報に基づき予測する。制御装置２０は、相互に異なる状態遷移情報（すなわち、各遷移情報ユニット）を用いることによって、複数の第２状態を予測する。説明の便宜上、第２状態と、予測された第２状態とを区別するため、予測された第２状態を「擬似状態」と表す。すなわち、制御装置２０は、相互に異なる状態遷移情報（すなわち、各遷移情報ユニット）を用いることによって、擬似状態を作成する。

[0040] 状態遷移情報を、ニューラルネットワークを用いて作成する場合に、制御装置２０は、第１状態、及び、当該第１状態における候補アクションを表す情報のうち、少なくともいずれかに対して当該状態遷移情報を適用することによって、該擬似状態を作成する。

[0041] 処理フェーズ３に関して上述した処理によって、制御装置２０は、各候補アクションに関して、複数の擬似状態を作成する。制御装置２０は、各候補アクションに関して、複数の擬似状態のばらつき度を算出する。

[0042] 制御装置２０は、複数の候補アクションの中から、該ばらつき度に基づきアクションを選択する。制御装置２０は、複数の候補アクションの中から、算出したばらつき度が上位である候補アクションを特定し、特定した候補アクションの中からアクションを選択する。制御装置２０は、たとえば、複数の候補アクションの中から、算出したばらつき度が最も大きな候補アクションを選択してもよい。

[0043] 上位は、たとえば、ばらつき度が最も大きいものから、ばらつき度が大きい順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。

[0044] 制御装置２０は、報酬情報を用いて、１つのアクション後の擬似状態における報酬程度を求め、求めた報酬程度と、当該１つのアクションに対するばらつき度とに基づき、アクションを選択してもよい。

[0045] 擬似状態が複数である場合に、制御装置２０は、たとえば、各擬似状態に関する報酬程度の平均（または、中央値）を求めることによって、アクションに関する報酬程度を求める。または、制御装置２０は、たとえば、各擬似状態の頻度が上位の状態を求め、求めた状態に関する報酬程度の平均（または、中央値）を求めることによって、アクションに関する報酬程度を求める。この場合に、上位は、たとえば、頻度が最も高いものから、頻度が高い順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。アクションに関する報酬程度を求める処理は、上述した例に限定されない。

[0046] また、１つのアクションに関する報酬程度と、該１つのアクションに関するばらつき度とに基づき、アクションを選択する処理は、たとえば、該報酬程度と、該ばらつき度とを足し算してもよいし、該報酬程度と、該ばらつき度との重み付き平均を算出してもよい。アクションを選択する処理は、上述した例に限定されない。

[0047] 制御装置２０は、アクションを選択した後に、選択したアクションを示す制御指令を指令実行装置５０へ出力する。指令実行装置５０は、受け取った制御指令が示すアクションを制御対象６０に関して実行する。

[0048] ＜制御装置の構成例＞

図２において制御装置２０は、演算装置３０と、記憶装置４０とを有している。演算装置３０は、状態推定部３１と、状態遷移情報更新部（状態遷移情報作成部）３２と、制御指令演算部３３と、予測状態決定部１１と、ばらつき度算出部１２と、候補アクション選択部１３とを有している。記憶装置

４０は、履歴情報記憶部４１と、状態遷移情報記憶部４２と、方策情報記憶部４３とを有している。

[0049] （処理フェーズ１）

状態推定部３１は、制御対象６０の第１状態を表す観測値（パラメタ値、センサー情報）を受け取る。状態推定部３１は、受け取ったセンサー情報と、状態遷移情報とに基づき、第１状態である制御対象６０に関してアクションを実行した後の制御対象６０の第２状態を推定する。状態推定部３１は、複数の候補アクションにおけるアクションに関して、それぞれ、第２状態を推定する処理を実行する。すなわち、状態推定部３１は、各候補アクションに関して擬似状態を作成する。

[0050] 制御指令演算部３３は、報酬情報を用いて、状態推定部３１によって作成された各擬似状態に関する報酬程度を算出する。制御指令演算部３３は、複数の候補アクションの中から、算出した報酬程度が上位の候補アクションのうちの１つアクションを選択する。制御指令演算部３３は、選択したアクションを示す制御指令を作成し、作成した制御指令を指令実行装置５０へ出力する。

[0051] 指令実行装置５０は、制御指令を受け取り、受け取った制御指令が示すアクションに従い、制御対象６０に関するアクションを実行する。制御対象６０に関するアクションの結果、制御対象６０の状態は、第１状態から第２状態に変化する。

[0052] 状態推定部３１は、制御対象６０の状態（この場合、第２状態）を表す観測値（パラメタ値、センサー情報）を受け取る。状態推定部３１は、第１状態と、第１状態にて実行されたアクションと、該第２状態とが関連付けされた履歴情報を作成し、作成した履歴情報を履歴情報記憶部４１に格納する。

[0053] 処理フェーズ１に関して上述したような処理を繰り返すことによって、履歴情報記憶部４１には、上記の履歴情報が蓄積される。

[0054] （処理フェーズ２）

説明の便宜上、ニューラルネットワーク等の統計的な手法（所定の処理手

順）を用いて状態遷移情報を作成する例を用いて、処理フェーズ２における処理を説明する。所定の処理手順は、例えば、ニューラルネット等の機械学習法に従った手順である。

[0055] 状態遷移情報更新部３２は、履歴情報記憶部４１に蓄積されている履歴情報を用いて、所定の処理手順に従って、複数の遷移情報ユニットを作成する。すなわち、状態遷移情報更新部３２は、該履歴情報を訓練データとして、所定の処理手順に従い状態遷移情報を作成し、作成した状態遷移情報を状態遷移情報記憶部４２に格納する。上述したように、状態遷移情報は、第１状態と、第２状態との関係性を表す。

[0056] 例えば、状態遷移情報更新部３２は、構成が互いに異なる複数のニューラルネットを用いて、複数の遷移情報ユニットを作成してもよい。構成が互いに異なる複数のニューラルネットは、例えば、互いにノードの数又はノード間の接続パタンが異なる複数のニューラルネットである。また、互いに構成の異なる複数のニューラルネットは、あるニューラルネットワークと、当該あるニューラルネットワークにおける一部のノードが存在していない（すなわち、一部のノードがドロップアウトしている）ニューラルネットワークとを用いて実現されていてもよい。

[0057] 状態遷移情報更新部３２は、パラメタの初期値が異なる複数のニューラルネットを用いて、複数の遷移情報ユニットを作成してもよい。

[0058] 状態遷移情報更新部３２は、履歴情報のうちの一部のデータ、または、履歴情報から重複を許してサンプリングしたものを訓練データとして用いてもよい。この場合に、複数の遷移情報ユニットは、相互に異なる訓練データに対して状態遷移情報を作成する。

[0059] なお、所定の処理手順は、ニューラルネットに限定されない。例えば所定の処理手順は、ＳＶＭ（ｓｕｐｐｏｒｔ ｖｅｃｔｏｒ ｍａｃｈｉｎｅ）、ランダムフォレスト、バギング（ｂｏｏｔｓｔｒａｐ ａｇｇｒｅｇａｔｉｎｇ）、又は、ベイジアンネットワークを算出する手順であってもよい。

[0060] （処理フェーズ３）

予測状態決定部１１は、複数の候補アクションについて、候補アクションをそれぞれ対象に関して施した後における第２状態を、状態遷移情報に基づき予測する。予測状態決定部１１は、相互に異なる状態遷移情報（すなわち、各遷移情報ユニット）を用いることによって、複数の擬似状態を作成する。

[0061] ばらつき度算出部１２は、予測状態決定部１１によって作成された複数の疑似状態のばらつき度（たとえば、分散値、エントロピー等）を算出し、算出したばらつき度を候補アクション選択部１３へ出力する。ばらつき度は、たとえば、分散値にある数を加えた値等であってもよく、上述した例に限定されない。

[0062] 候補アクション選択部１３は、複数の候補アクションの中から、該ばらつき度に基づきアクションを選択する。候補アクション選択部１３は、複数の候補アクションの中から、算出したばらつき度が上位である候補アクションを特定し、特定した候補アクションの中からアクションを選択する。候補アクション選択部１３は、たとえば、複数の候補アクションの中から、算出したばらつき度が最も大きな候補アクションを選択してもよい。

[0063] 制御指令演算部３３は、候補アクション選択部１３が選択したアクションを示す制御指令を作成し、作成した制御指令を指令実行装置５０へ出力する。

[0064] 上述したように候補アクション選択部１３は、ばらつき度が大きいアクションを選択する。ばらつき度は、状態遷移情報に従い算出された結果がばらついていることを表している。このため、ばらつき度が大きい場合には、状態遷移情報が不安定であることを表しているということもできる。すなわち、ばらつき度が大きいアクションを実行することによって、探索が十分でない状態遷移を積極的に探索することができるという効果を奏する。

[0065] 候補アクション選択部１３は、状態価値情報に基づき、状態に関する価値の程度を表す状態価値情報を作成してもよい。状態価値情報は、たとえば、状態に対して、当該状態の価値の程度を表す関数である。この場合に、価値

は、当該状態を実現することが望ましい程度を表す情報であるとも言うことができる。状態価値情報は、アクション後における制御対象６０の状態がどの程度望ましいのかを表す情報ともいうことができる。状態価値情報は、また、当該アクションがどの程度の望ましいのかを表す情報ともいうことができる。

[0066] 候補アクション選択部１３は、状態価値情報を作成する処理において、報酬情報を用いてもよい。たとえば、候補アクション選択部１３は、各アクションに関して算出されたばらつき度を、新たに、状態価値情報として設定してもよい。たとえば、候補アクション選択部１３は、各アクションに関して算出されたばらつき度を状態価値情報として設定し、その後、当該アクションに関する報酬情報を加える等の処理を実行することによって、状態価値情報を更新してもよい。この場合に、ばらつき度は、報酬情報に対する追加的な報酬（疑似追加報酬）であるともいうことができる。

[0067] 状態価値情報を作成する処理は、上述した例に限定されず、たとえば、報酬情報に所定の値を加算した値、報酬情報に所定の値を減算した値、または、報酬情報に所定の値を乗算した値等に基づき実行されてもよい。すなわち、ばらつき度が大きいほど、状態価値情報は、価値の程度が高いことを表す情報であればよい。

[0068] 候補アクション選択部１３は、状態価値情報に基づき、複数の候補アクションの中から、該価値の程度が上位の候補アクションを選択し、選択した候補アクションからアクションを選択してもよい。候補アクション選択部１３は、たとえば、算出した価値が最も高い候補アクションを選択してもよい。この場合に、上位は、たとえば、価値の程度が最も高いものから価値の程度が高い順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。

[0069] 制御指令が作成された後に、指令実行装置５０は、該制御指令を受け取り、受け取った制御指令が示すアクションに従い、制御対象６０に関するアクションを実行する。制御対象６０に関するアクションの結果、制御対象６０

の状態は、第１状態から第２状態に変化する。

[0070] 状態推定部３１は、制御対象６０の状態（この場合、第２状態）を表す観測値（パラメタ値、センサー情報）を受け取る。状態推定部３１は、第１状態と、第１状態にて実行されたアクションと、該第２状態とが関連付けされた履歴情報を作成し、作成した履歴情報を履歴情報記憶部４１に格納する。

[0071] 処理フェーズ３に関して上述したような処理が、複数のタイミングに関して実行されることにより、履歴情報記憶部（不図示）には、複数のタイミングにおける履歴情報が蓄積される。

[0072] ＜制御装置の動作例＞

以上の構成を有する演算装置３０の処理動作の一例について説明する。図３は、第２実施形態の演算装置の処理動作の一例を示すフローチャートである。図３に示すフローチャートにおいて、ステップＳ１０１は、上記の処理フェーズ１に対応し、ステップＳ１０２は、処理フェーズ２に対応し、ステップＳ１０３，Ｓ１０４は、処理フェーズ３に対応する。

[0073] 演算装置３０は、履歴情報が蓄積されるまで、処理フェーズ１及び処理フェーズ２、または、処理フェーズ３及び処理フェーズ２のうち、少なくとも、いずれかの処理を繰り返すことによって、履歴情報を取得する（ステップＳ１０１）。

[0074] 演算装置３０は、処理フェーズ２に示された処理に従い、状態遷移情報を更新する（ステップＳ１０２）。

[0075] 演算装置３０は、処理フェーズ３にて上述した処理に従い、ばらつき度を算出する（ステップＳ１０３）。

[0076] 演算装置３０は、履歴情報に基づき方策情報を更新する（ステップＳ１０４）。具体的には、演算装置３０は、履歴情報に基づき、第１状態と、当該第１状態にて実行したアクションと、第２状態と特定し、特定したこれらの情報を用いて、方策情報を更新する。そして、処理ステップは、ステップＳ１０１（処理フェーズ１）に戻る。

[0077] なお、以上の説明では、演算装置３０が、処理フェーズ３にて、履歴情報

を蓄積してから方策情報を更新し、その後直ぐに、処理フェーズ１に戻るものとして説明した。説明の便宜上、本実施形態においては、図３を参照しながら上述した処理を、「バッチ学習」と記載する。

すなわち、バッチ学習は、ある程度（説明の便宜上、「第１蓄積程度」と称する）の履歴情報が蓄積されてから、該履歴情報を用いて方策情報を更新（または、作成）する処理を表す。第１蓄積程度は、履歴が複数であることを表している。ただし、演算装置３０における処理は、上述したバッチ学習に限定されず、例えば、方策情報は、オンライン学習によって更新（または、作成）されてもよい、ミニバッチ学習によって更新（または、作成）されてもよい。

[0078] オンライン学習は、履歴情報に履歴が１つ追加されるごとに、該履歴情報を用いて方策情報を更新（または、作成）する処理を表す。

[0079] ミニバッチ学習は、ある程度（説明の便宜上、「第２蓄積程度」と称する）の履歴情報が蓄積されてから、該履歴情報を用いて方策情報を更新（または、作成）する処理を表す。第２蓄積程度は、履歴が複数であることを表している。ミニバッチ学習は、バッチ学習と類似した処理である。しかし、第２蓄積程度は、第１蓄積程度に比べて少ない。

[0080] 第１蓄積程度、及び、第２蓄積程度は、必ずしも、処理フェーズ１乃至処理フェーズ３に示された反復処理ごとに一定の程度でなくともよく、該反復処理ごと異なる個数を表していてもよい。

[0081] オンライン学習の場合、履歴情報を取得する度に方策情報を更新して、ステップＳ１０１（処理フェーズ１）へ戻るように、修正されてもよい。すなわち、オンライン学習の場合には、候補アクション選択部１３は、第２状態に関するセンサー情報が届く度に、ポリシーモデルを更新する。

[0082] 「ミニバッチ学習」は、方策情報の更新タイミング以外は、上記「オンライン学習」の処理動作と変わらない。すなわち、「ミニバッチ学習」にて一度の方策情報の更新に用いられる履歴情報量は、「オンライン学習」よりも多いので、「ミニバッチ学習」における方策情報の更新周期は、「オンライ

ン学習」よりも長くなる。

[0083] ＜第３実施形態＞

第３実施形態は、より具体的な実施形態に関する。すなわち、第３実施形態は、第２実施形態のバリエーションに関する。

[0084] ＜制御装置の概要＞

図４は、第３実施形態の演算装置８０を含む制御装置７０の一例を示すブロック図である。図４には、制御装置７０の他に、図２と同様に指令実行装置５０及び制御対象６０が図示されている。

[0085] 制御装置７０は、後述するような、「処理フェーズ１」、「処理フェーズ２」、及び「処理フェーズ３」を実行する。制御装置７０は、これらの処理を実行することによって、制御対象６０の状態（ｓｔａｔｅ）が、より早期に所望状態に近付くよう、方策情報を学習する。

[0086] 方策情報は、制御対象６０がある状態である場合に、実行可能なアクションを表す。方策情報は、たとえば、該ある状態と、該アクションとが関連付けされた情報を用いて実現することができる。方策情報は、たとえば、該ある状態を与えた場合に該アクションを算出する処理であってもよい。当該処理は、たとえば、ある関数、または、統計的な手法によって算出された、当該ある状態と、当該アクションとの関係性を表すモデルであってもよい。すなわち、方策情報は、上述した例に限定されない。

[0087] 以降の説明においては、説明の便宜上、制御対象６０は、車両や、発電機等（以降、「実環境」と表す）であるとする。あるタイミング（以降、「第１タイミング」と表す）における制御対象６０の状態を「第１状態」と表す。あるタイミングの次のタイミング（以降、「第２タイミング」と表す）における制御対象６０の状態を「第２状態」と表す。制御対象６０の状態は、第１状態に応じたアクションが実施された後に第２状態に変化するとする。また、第１状態と、第２状態とは、必ずしも、相互に異なる状態である必要はなく、同じ状態を表していてもよい。以降の説明においては、説明の便宜上、第１状態と、第２状態との異同によらずに、「第１状態から第２状態に

変化する」と表すこととする。

[0088] 制御装置７０は、後述する「処理フェーズ１」にて、複数のタイミングに関して、制御対象６０の状態を参照しながら後述するような処理を実行することによって、タイミングごとにアクションを決定する。すなわち、制御装置７０は、第１タイミングに関して処理を実行した後に、第２タイミングに関して処理を実行し、さらに、第２タイミングより後のタイミングに関しても処理を実行する。したがって、第１タイミング、及び、第２タイミングは、特定のタイミングを表しているわけではなく、制御装置７０における処理に関して連続している２つのタイミングを表している。

[0089] （処理フェーズ１）

制御装置７０は、第１状態である制御対象６０に関して、第１状態と方策情報とに基づきアクションを決定し、決定したアクションを示す制御指令を、指令実行装置５０へ出力する。

[0090] 指令実行装置５０は、制御装置７０によって制御指令を受け取り、受け取った該制御指令が示すアクションを、制御対象６０に関して実行する。この結果、制御対象６０の状態は、第１状態から第２状態に変化する。

[0091] 説明の便宜上、制御対象６０には、制御対象６０を観測しているセンサー（図示せず）が取り付けられているとする。センサーは、制御対象６０に関して観測した観測値を表すセンサー情報を作成し、作成したセンサー情報を出力するとする。制御対象６０を観測しているセンサーは、複数であってもよい。

[0092] 制御装置７０は、第１状態に関するアクションが実行された後に、センサーによって作成された該センサー情報を受け取り、受け取った該センサー情報に関する第２状態を推定する。制御装置７０は、該第１状態と、該アクションと、該第２状態とが関連付けされた情報（以降、「履歴情報」と表す）を作成する。制御装置７０は、作成した履歴情報を、後述する履歴情報記憶部９１に格納してもよい。

[0093] 処理フェーズ１に関して上述したような処理が、複数のタイミングに関し

て実行されることにより、後述する履歴情報記憶部４１には、複数のタイミングにおける履歴情報が蓄積される。

[0094] （処理フェーズ２）

制御装置７０は、処理フェーズ１にて蓄積された履歴情報を用いて、状態遷移情報を更新する（または、作成する）。状態遷移情報を、ニューラルネットワークを用いて作成する場合に、制御装置７０は、上述したような履歴情報に含まれているデータを訓練データとして用いて、当該状態遷移情報を作成する。後述するように、制御装置７０は、たとえば、構成が相互に異なっているニューラルネットワークを用いて、複数の状態遷移情報を作成する。

[0095] ここで、状態遷移情報について説明する。状態遷移情報は、第１状態と、第２状態との間の関係性を表す情報であり、たとえば、制御対象６０の状態遷移（つまり、アクションによる第１状態から第２状態への状態遷移）を、履歴情報を用いてモデル化したものである。すなわち、状態遷移情報を用いることにより、第１状態とアクションとの組み合わせに対応する第２状態を予測することができる。以降、制御対象６０の第１状態及び第２状態と区別するために、状態遷移情報の第１状態及び第２状態を、「第１疑似状態」及び「第２疑似状態」と表すことがある。また、「第２疑似状態」を「予測状態」と表すことがある。

[0096] （処理フェーズ３）

制御装置７０は、状態遷移情報に基づき、第１疑似状態にて可能な複数の「候補アクション」のそれぞれに関して複数の「予測状態」を決定する。制御装置７０は、相互に異なる状態遷移情報（すなわち、各遷移情報ユニット）を用いることによって、複数の第２疑似状態を作成する。

[0097] 状態遷移情報を、ニューラルネットワークを用いて作成する場合に、制御装置７０は、第１疑似状態、及び、当該第１疑似状態における候補アクションを表す情報に対して当該状態遷移情報を適用することによって、第２擬似状態を作成する。

[0098] 処理フェーズ３に関して上述した処理によって、制御装置７０は、各候補アクションに関して、複数の予測状態を作成する。制御装置７０は、各候補アクションに関して、複数の予測状態のばらつき度を算出する。

[0099] 制御装置７０は、複数の候補アクションの中から、該ばらつき度に基づきアクションを選択する。この選択されたアクションは、後述するように、方策情報の更新に用いられるので、以降、「更新使用アクション」と表すことがある。制御装置７０は、複数の候補アクションの中から、算出したばらつき度が上位である候補アクションを特定し、特定した候補アクションの中から更新使用アクションを選択する。制御装置７０は、たとえば、複数の候補アクションの中から、算出したばらつき度が最も大きな候補アクションを選択してもよい。

[0100] 上位は、たとえば、ばらつき度が最も大きいものから、ばらつき度が大きい順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。

[0101] 制御装置７０は、報酬情報を用いて、１つの候補アクション後の予測状態における報酬程度を求め、求めた報酬程度と、当該１つの候補アクションに対するばらつき度とに基づき、更新使用アクションを選択してもよい。報酬情報は、ある状態が望ましい程度（つまり、「報酬程度」）を表す。報酬情報は、たとえば、該ある状態と、該程度とが関連付けされた情報を用いて実現することができる。報酬情報は、たとえば、該ある状態を与えた場合に該報酬程度を算出する処理であってもよい。当該処理は、たとえば、ある関数、または、統計的な手法によって算出された、当該ある状態と、当該報酬程度との関係性を表すモデルであってもよい。すなわち、報酬情報は、上述した例に限定されない。

[0102] 予測状態が複数である場合に、制御装置７０は、たとえば、各予測状態に関する報酬程度の平均（または、中央値）を求めることによって、候補アクションに関する報酬程度を求める。または、制御装置７０は、たとえば、各予測状態の頻度が上位の状態を求め、求めた状態に関する報酬程度の平均（

または、中央値）を求めることによって、候補アクションに関する報酬程度を求める。この場合に、上位は、たとえば、頻度が最も高いものから、頻度が高い順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。候補アクションに関する報酬程度を求める処理は、上述した例に限定されない。

[0103] また、１つの候補アクションに関する報酬程度と、該１つの候補アクションに関するばらつき度とに基づき、更新使用アクションを選択する処理は、たとえば、該報酬程度と、該ばらつき度とを足し算してもよいし、該報酬程度と、該ばらつき度との重み付き平均を算出してもよい。更新使用アクションを選択する処理は、上述した例に限定されない。

[0104] 制御装置７０は、更新使用アクションに基づき、方策情報を更新する。たとえば、制御装置７０は、更新使用アクションが処理フェーズ１にて確定的に又は他のアクションに比べて高い確率で選択されるように、方策情報を更新する。この更新された方策情報は、処理フェーズ１にて用いられることになる。

[0105] ＜制御装置の構成例＞

図４において制御装置７０は、演算装置８０と、記憶装置９０とを有している。演算装置３０は、状態推定部８１と、状態遷移情報更新部（状態遷移情報作成部）８２と、制御指令演算部８３と、予測状態決定部１１と、ばらつき度算出部１２と、候補アクション選択部１３とを有している。記憶装置９０は、履歴情報記憶部９１と、状態遷移情報記憶部９２と、方策情報記憶部９３とを有している。以降、制御装置７０の構成を処理フェーズ毎に説明する。

[0106] （処理フェーズ１）

状態推定部８１は、制御対象６０の状態を表す観測値（パラメタ値、センサー情報）を受け取る。状態推定部８１は、受け取った観測値（パラメタ値、センサー情報）に基づき、制御対象６０の状態を推定する。

[0107] 制御指令演算部８３は、状態推定部８１に推定された状態と方策情報記憶

部９３に記憶されている方策情報とに基づきアクションを決定し、決定したアクションを示す制御指令を、指令実行装置５０へ出力する。指令実行装置５０は、制御装置７０によって制御指令を受け取り、受け取った該制御指令が示すアクションを、制御対象６０に関して実行する。この結果、制御対象６０の状態は、第１状態から第２状態に変化する。

[0108] 状態推定部８１は、制御対象６０の状態（この場合、第２状態）を表す観測値（パラメタ値、センサー情報）を受け取る。状態推定部８１は、第１状態と、第１状態にて実行されたアクションと、該第２状態とが関連付けされた履歴情報を作成し、作成した履歴情報を履歴情報記憶部９１に格納する。

[0109] 処理フェーズ１に関して上述したような処理を繰り返すことによって、履歴情報記憶部９１には、上記の履歴情報が蓄積される。

[0110] （処理フェーズ２）

説明の便宜上、ニューラルネットワーク等の統計的な手法（所定の処理手順）を用いて状態遷移情報を作成する例を用いて、処理フェーズ２に対応する制御装置７０の構成について説明する。所定の処理手順は、例えば、ニューラルネット等の機械学習法に従った手順である。

[0111] 状態遷移情報更新部８２は、履歴情報記憶部９１に蓄積されている履歴情報を用いて、所定の処理手順に従って、複数の状態遷移情報を作成する。すなわち、状態遷移情報更新部８２は、該履歴情報を訓練データとして、所定の処理手順に従い状態遷移情報を作成し、作成した状態遷移情報を状態遷移情報記憶部９２に格納する。上述したように、状態遷移情報は、第１状態と、第２状態との関係性を表す。

[0112] たとえば、状態遷移情報更新部８２は、構成が互いに異なる複数のニューラルネットを用いて、複数の遷移情報ユニットを作成してもよい。構成が互いに異なる複数のニューラルネットは、例えば、互いにノードの数又はノード間の接続パタンが異なる複数のニューラルネットである。また、互いに構成の異なる複数のニューラルネットは、あるニューラルネットワークと、当該あるニューラルネットワークにおける一部のノードが存在していない（す

なわち、一部のノードがドロップアウトしている）ニューラルネットワークとを用いて実現されていてもよい。

[0113] 状態遷移情報更新部８２は、パラメタの初期値が異なる複数のニューラルネットを用いて、複数の遷移情報ユニットを作成してもよい。

[0114] 状態遷移情報更新部８２は、履歴情報のうちの一部のデータ、または、履歴情報から重複を許してサンプリングしたものを訓練データとして用いてもよい。この場合に、複数の遷移情報ユニットは、相互に異なる訓練データに対して状態遷移情報を作成する。

[0115] なお、所定の処理手順は、ニューラルネットに限定されない。例えば所定の処理手順は、ＳＶＭ（ｓｕｐｐｏｒｔ ｖｅｃｔｏｒ ｍａｃｈｉｎｅ）、ランダムフォレスト、バギング（ｂｏｏｔｓｔｒａｐ ａｇｇｒｅｇａｔｉｎｇ）、又は、ベイジアンネットワークを算出する手順であってもよい。

[0116] （処理フェーズ３）

制御指令演算部８３は、第１疑似状態にて可能な複数の候補アクションをそれぞれ示す複数の制御指令を予測状態決定部１１へ出力する。

[0117] 予測状態決定部１１は、第１疑似状態にて可能な複数の候補アクションと状態遷移情報とに基づき、第１疑似状態にて可能な複数の「候補アクション」のそれぞれに関して複数の予測状態を決定する。制御装置７０は、相互に異なる状態遷移情報（すなわち、各遷移情報ユニット）を用いることによって、各候補アクションに関して複数の第２疑似状態を作成する。

[0118] 制御指令演算部８３は、予測状態決定部１１にて作成された各第２疑似状態を新たな第１疑似状態として、該第１疑似状態にて可能な複数の候補アクションをそれぞれ示す複数の制御指令を予測状態決定部１１へ出力する。このとき、制御指令演算部８３は、たとえば、予測状態決定部１１にて複数の状態遷移情報のうちのある１つを用いて作成された各第２状態情報を新たな第１疑似状態としてもよい。

[0119] 上述したような制御指令演算部８３と予測状態決定部１１との遣り取りによって、候補アクション選択部１３には、第１疑似状態、第２疑似状態、及

び候補アクションの各組み合わせに対応する、ばらつき度が蓄積されることになる。

[0120] ばらつき度算出部１２は、予測状態決定部１１によって作成された複数の予測状態のばらつき度（たとえば、分散値、エントロピー等）を算出し、算出したばらつき度を候補アクション選択部１３へ出力する。ばらつき度は、たとえば、分散値にある数を加えた値等であってもよく、上述した例に限定されない。

[0121] 候補アクション選択部１３は、複数の候補アクションの中から、該ばらつき度に基づき更新使用アクションを選択する。候補アクション選択部１３は、たとえば、複数の候補アクションの中から、算出したばらつき度が上位である候補アクションを特定し、特定した候補アクションの中から更新使用アクションを選択する。候補アクション選択部１３は、たとえば、複数の候補アクションの中から、算出したばらつき度が最も大きな候補アクションを選択してもよい。

[0122] 候補アクション選択部１３は、更新使用アクションに基づき、方策情報を更新する。たとえば、候補アクション選択部１３は、処理フェーズ１にて制御指令演算部８３によって更新使用アクションが確定的に又は他の候補アクションに比べて高い確率で選択されるように、方策情報記憶部９３に記憶されている方策情報を更新する。

[0123] 上述したように候補アクション選択部１３は、ばらつき度が大きい候補アクションを選択する。ばらつき度は、状態遷移情報に従い算出された結果がばらついていることを表している。このため、ばらつき度が大きい場合には、状態遷移情報が不安定であることを表しているということもできる。すなわち、ばらつき度が大きいアクションを実行することによって、探索が十分でない状態遷移を積極的に探索することができるという効果を奏する。

[0124] 候補アクション選択部１３は、状態価値情報に基づき、状態に関する価値の程度を表す状態価値情報を作成してもよい。状態価値情報は、たとえば、状態に対して、当該状態の価値の程度を表す関数である。この場合に、価値

は、当該状態を実現することが望ましい程度を表す情報であるとも言うことができる。状態価値情報は、アクション後における制御対象６０の状態がどの程度望ましいのかを表す情報ともいうことができる。状態価値情報は、また、当該アクションがどの程度の望ましいのかを表す情報ともいうことができる。

[0125] 候補アクション選択部１３は、状態価値情報を作成する処理において、報酬情報を用いてもよい。たとえば、候補アクション選択部１３は、各候補アクションに関して算出されたばらつき度を、新たに、状態価値情報として設定してもよい。たとえば、候補アクション選択部１３は、各候補アクションに関して算出されたばらつき度を状態価値情報として設定し、その後、当該候補アクションに関する報酬情報を加える等の処理を実行することによって、状態価値情報を更新してもよい。この場合に、ばらつき度は、報酬情報に対する追加的な報酬（疑似追加報酬）であるともいうことができる。

[0126] 状態価値情報を作成する処理は、上述した例に限定されず、たとえば、報酬情報に所定の値を加算した値、報酬情報に所定の値を減算した値、または、報酬情報に所定の値を乗算した値等に基づき実行されてもよい。すなわち、ばらつき度が大きいほど、状態価値情報は、価値が高いことを表す情報であればよい。

[0127] 候補アクション選択部１３は、状態価値情報に基づき、複数の候補アクションの中から、該価値の程度が上位の候補アクションを選択し、選択した候補アクションから更新使用アクションを選択してもよい。候補アクション選択部１３は、たとえば、算出した価値が最も高い候補アクションを選択してもよい。この場合に、上位は、たとえば、価値の程度が最も高いものから価値の程度が高い順に数えて、１％、５％、または、１０％等の所定の割合以内であることを表している。

[0128] ＜制御装置の動作例＞

以上の構成を有する演算装置８０の処理動作の一例について説明する。図５は、第３実施形態の演算装置の処理動作の一例を示すフローチャートであ

る。図５に示すフローチャートにおいて、ステップＳ２０１は、上記の処理フェーズ１に対応し、ステップＳ２０２は、処理フェーズ２に対応し、ステップＳ２０３，Ｓ２０４は、処理フェーズ３に対応する。

[0129] 演算装置８０は、履歴情報が蓄積されるまで、処理フェーズ１に示された処理を繰り返すことによって、履歴情報を取得する（ステップＳ２０１）。

[0130] 演算装置８０は、処理フェーズ２に示された処理によって、状態遷移情報を更新する（ステップＳ２０２）。

[0131] 演算装置８０は、ばらつき度が蓄積されるまで、処理フェーズ３に示された処理によって、ばらつき度を算出する（ステップＳ２０３）。

[0132] 演算装置８０は、ばらつき度に基づき方策情報を更新する（ステップＳ２０４）。そして、処理ステップは、ステップＳ２０１（処理フェーズ１）に戻る。

[0133] なお、以上の説明では、演算装置８０が、処理フェーズ３にて、ばらつき度を蓄積してから方策情報を更新し、その後直ぐに、処理フェーズ１に戻るものとして説明した。すなわち、以上の説明では、方策情報がバッチ学習されるケースを例にとり説明したが、これに限定されるものではない。例えば、方策情報は、オンライン学習されてもよいし、ミニバッチ学習されてもよい。

[0134] 「オンライン学習」の場合、図５のフローチャートは、ステップＳ２０３，Ｓ２０４の処理を繰り返しループとし、該ループが所定回数繰り返されたことを条件に、ステップＳ２０１（処理フェーズ１）へ戻るように、修正されてもよい。すなわち、「オンライン学習」の場合、候補アクション選択部１３は、ばらつき度が届く度に、方策情報を更新することになる。

[0135] 「ミニバッチ学習」の場合、図５のフローチャートは、「オンライン学習」の場合と同様に、ステップＳ２０３，Ｓ２０４の処理を繰り返しループとし、該ループが所定回数繰り返されたことを条件に、ステップＳ２０１（処理フェーズ１）へ戻るように、修正されてもよい。ただし、「ミニバッチ学習」の場合、候補アクション選択部１３は、「オンライン学習」の場合と異

なり、複数個のばらつき度が蓄積されたタイミングで、方策情報を更新することになる。

[0136] ＜他の実施形態＞

図６は、演算装置のハードウェア構成例を示す図である。図６において演算装置１００は、プロセッサ１０１と、メモリ１０２とを含んでいる。第１実施形態及び第２実施形態で説明した演算装置１０，３０，８０の状態推定部３１，８１と、状態遷移情報更新部（状態遷移情報作成部）３２，８２と、制御指令演算部３３，８３と、予測状態決定部１１と、ばらつき度算出部１２と、候補アクション選択部１３とは、プロセッサ１０１がメモリ１０２に記憶されたプログラムを読み込んで実行することにより実現されてもよい。プログラムは、様々なタイプの非一時的なコンピュータ可読媒体（ｎｏｎ－ｔｒａｎｓｉｔｏｒｙ　ｃｏｍｐｕｔｅｒ　ｒｅａｄａｂｌｅ　ｍｅｄｉｕｍ）を用いて格納され、演算装置１０，３０，８０に供給することができる。また、プログラムは、様々なタイプの一時的なコンピュータ可読媒体（ｔｒａｎｓｉｔｏｒｙ　ｃｏｍｐｕｔｅｒ　ｒｅａｄａｂｌｅ　ｍｅｄｉｕｍ）によって演算装置１０，３０，８０に供給されてもよい。

[0137] 上述したような演算装置は、たとえば、製造工場における各装置を制御する制御装置としても機能することができる。この場合に、各製造工場には、各装置や、製造工場における状態（たとえば、気温、湿度、視界等）等を測定するセンサーが配置される。各センサーは、各装置や、製造工場における状態等を測定し、測定した状態を表す観測情報を作成する。この場合に、観測情報は、製造工場において観測される状態を表す情報である。

[0138] 演算装置は、当該観測情報を受け取り、上述したような処理を行うことにより決定されたアクションに従い、各装置を制御する。たとえば、装置が、材料の量を調整するバルブである場合に、演算装置は、決定したアクションに従い、バルブを閉める、または、バルブを開ける等の制御を行う。または、装置が、温度を調整するヒータである場合に、演算装置は、決定したアクションに従い、設定温度を上げる、または、設定温度を下げる等の制御を行

う。

[0139] 製造工場における各装置を制御する例を参照しながら、制御例について説明したが、制御例は、上述した例に限定されない。たとえば、演算装置は、上述したような処理と同様な処理を行うことによって、化学工場における各装置を制御する制御装置、発電所における各装置を制御する制御装置としても機能することができる。

[0140] 以上、実施の形態を参照して本願発明を説明したが、本願発明は上記によって限定されるものではない。本願発明の構成や詳細には、発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

符号の説明

[0141] １０，３０，８０ 演算装置（アクション決定装置）

１１ 予測状態決定部

１２ ばらつき度算出部

１３ 候補アクション選択部

２０，７０ 制御装置

３１，８１ 状態推定部

３２，８２ 状態遷移情報更新部（状態遷移情報作成部）

３３，８３ 制御指令演算部

４０，９０ 記憶装置

４１，９１ 履歴情報記憶部

４２，９２ 状態遷移情報記憶部

４３，９３ 方策情報記憶部

５０ 指令実行装置

６０ 制御対象

請求の範囲

[請求項1] 第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定する決定手段と、

各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出する算出手段と、

前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する選択手段と、

を具備する演算装置。

[請求項2] 前記選択手段は、前記複数の候補アクションの中から、前記一部の候補アクションとして、前記ばらつき度が上位の前記候補アクションを選択する、

請求項１に記載の演算装置。

[請求項3] 前記選択手段は、前記一部の候補アクションとして、前記ばらつき度が最大の前記候補アクションを選択する、

請求項１に記載の演算装置。

[請求項4] ２つの状態と、該２つの状態間におけるアクションとが関連付けされたセットを含む履歴情報に基づき、所定の処理手順に従い、前記遷移情報を作成する作成手段をさらに具備する、

請求項１乃至請求項３のいずれかに記載の演算装置。

[請求項5] 前記所定の処理手順は、ニューラルネットを算出する手順である、

請求項４に記載の演算装置。

[請求項6] 前記作成手段は、複数の前記遷移情報を、互いに構成の異なる複数の前記ニューラルネットを用いて作成する、

請求項５に記載の演算装置。

[請求項7] 前記作成手段は、複数の前記遷移情報を、パラメタの初期値が異なる複数の前記ニューラルネットを用いて作成する、

請求項５に記載の演算装置。

[請求項8] 複数の前記遷移情報を、前記履歴情報のうちの互いに異なるセットを複数の前記ニューラルネットに入力することによって作成する、

請求項５に記載の演算装置。

[請求項9] 情報処理装置によって、第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定し、

各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出し、

前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する、

アクション決定方法。

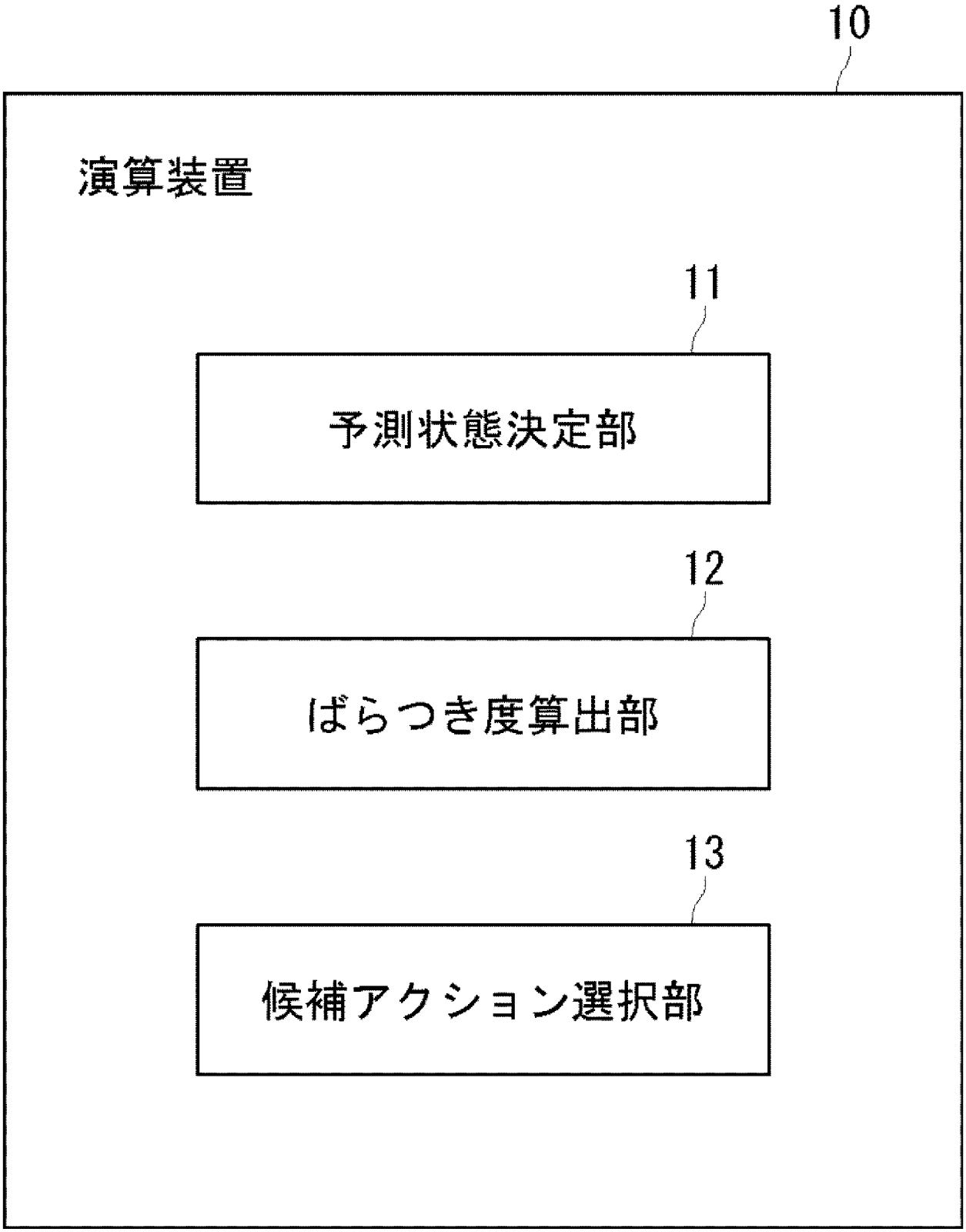
[請求項10] 第１タイミングでの第１状態と、前記第１タイミング以降の第２タイミングでの第２状態との関係性を表す遷移情報を複数用いて、第１状態にて可能な複数の候補アクションのそれぞれに関して複数の第２状態を決定し、

各前記候補アクションに関して、前記複数の第２状態のばらつき度を算出し、

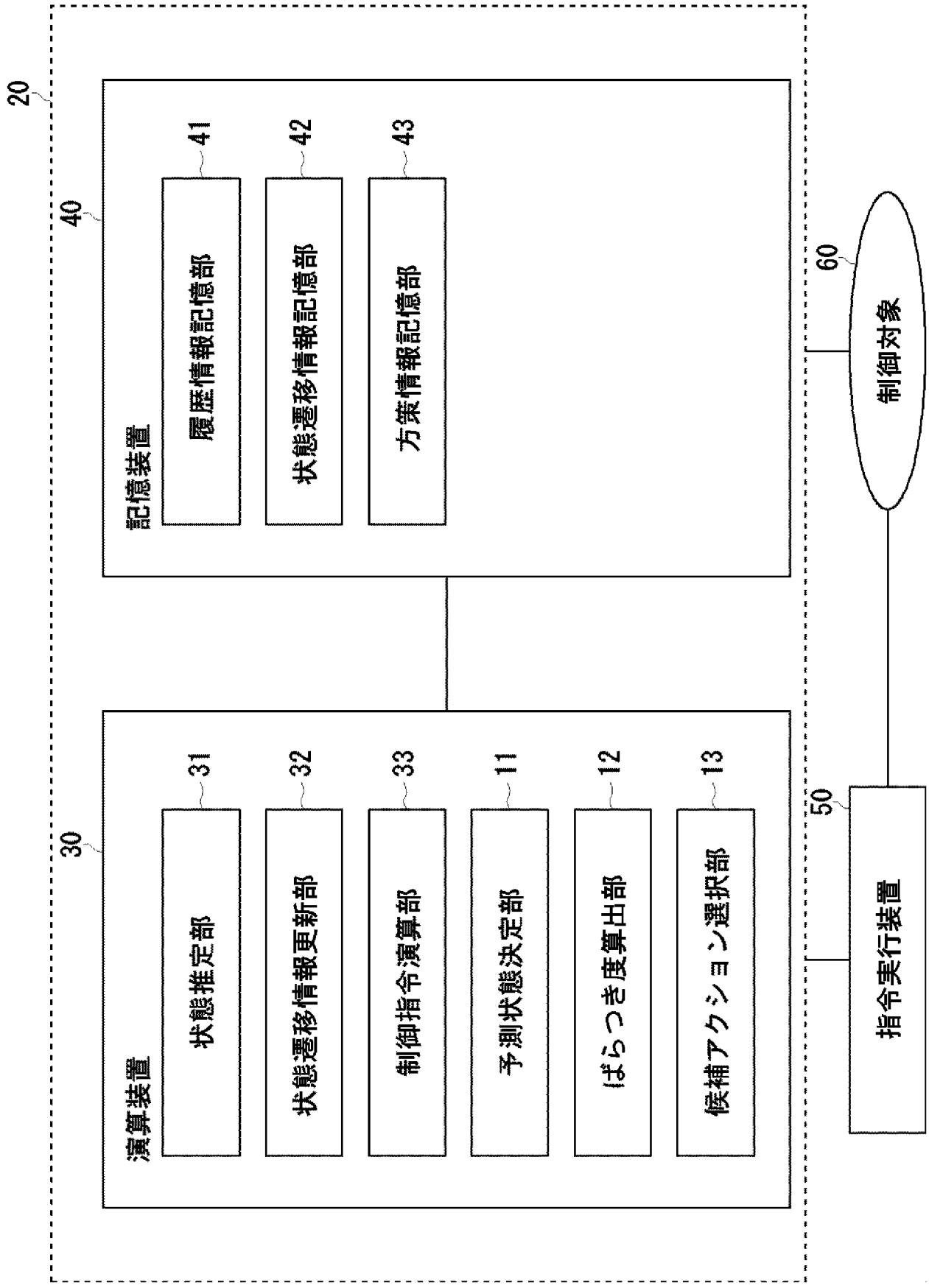
前記ばらつき度に基づき、前記複数の候補アクションから一部の候補アクションを選択する、

処理を、演算装置に実行させる制御プログラムを格納する非一時的なコンピュータ可読媒体。

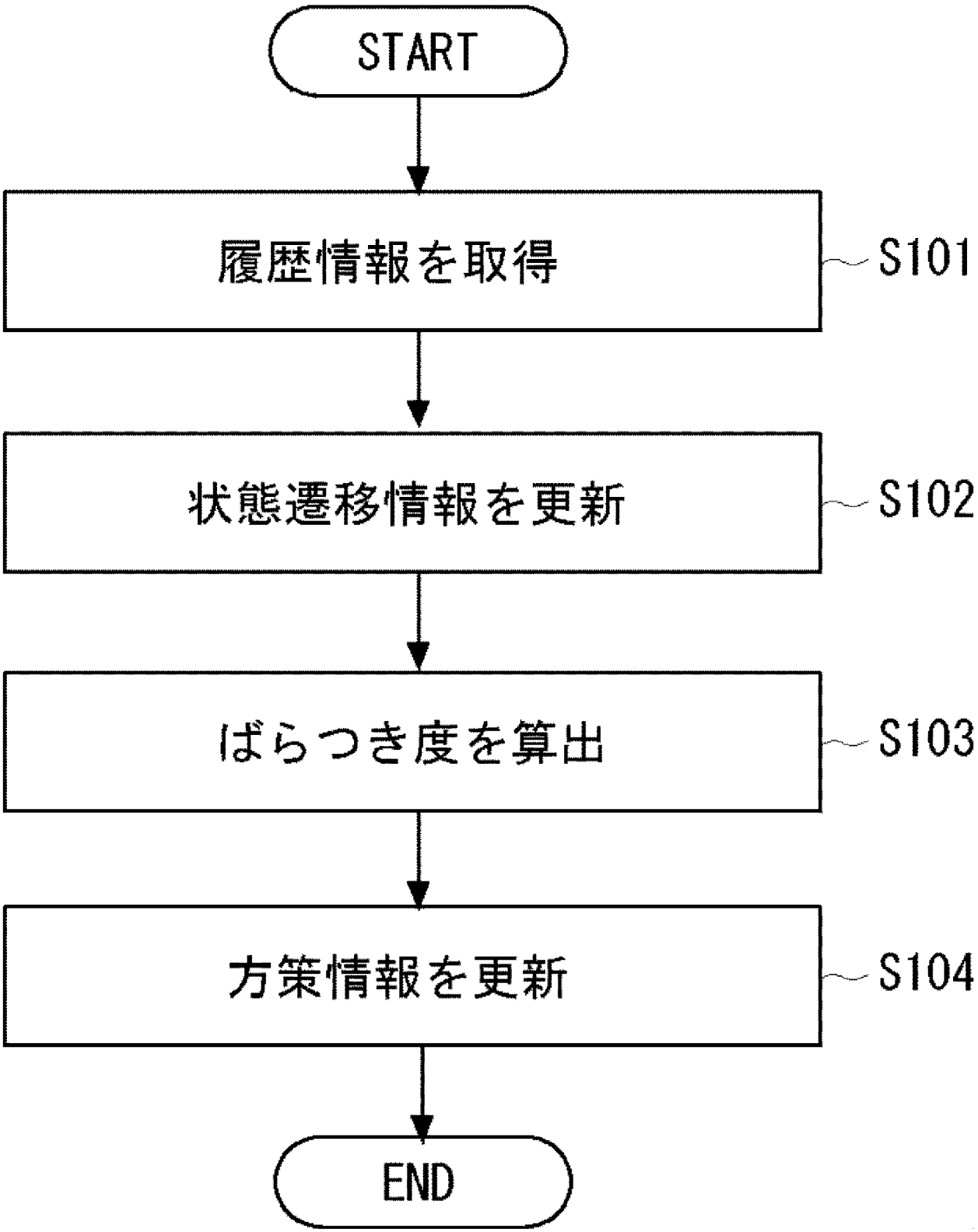
[図1]



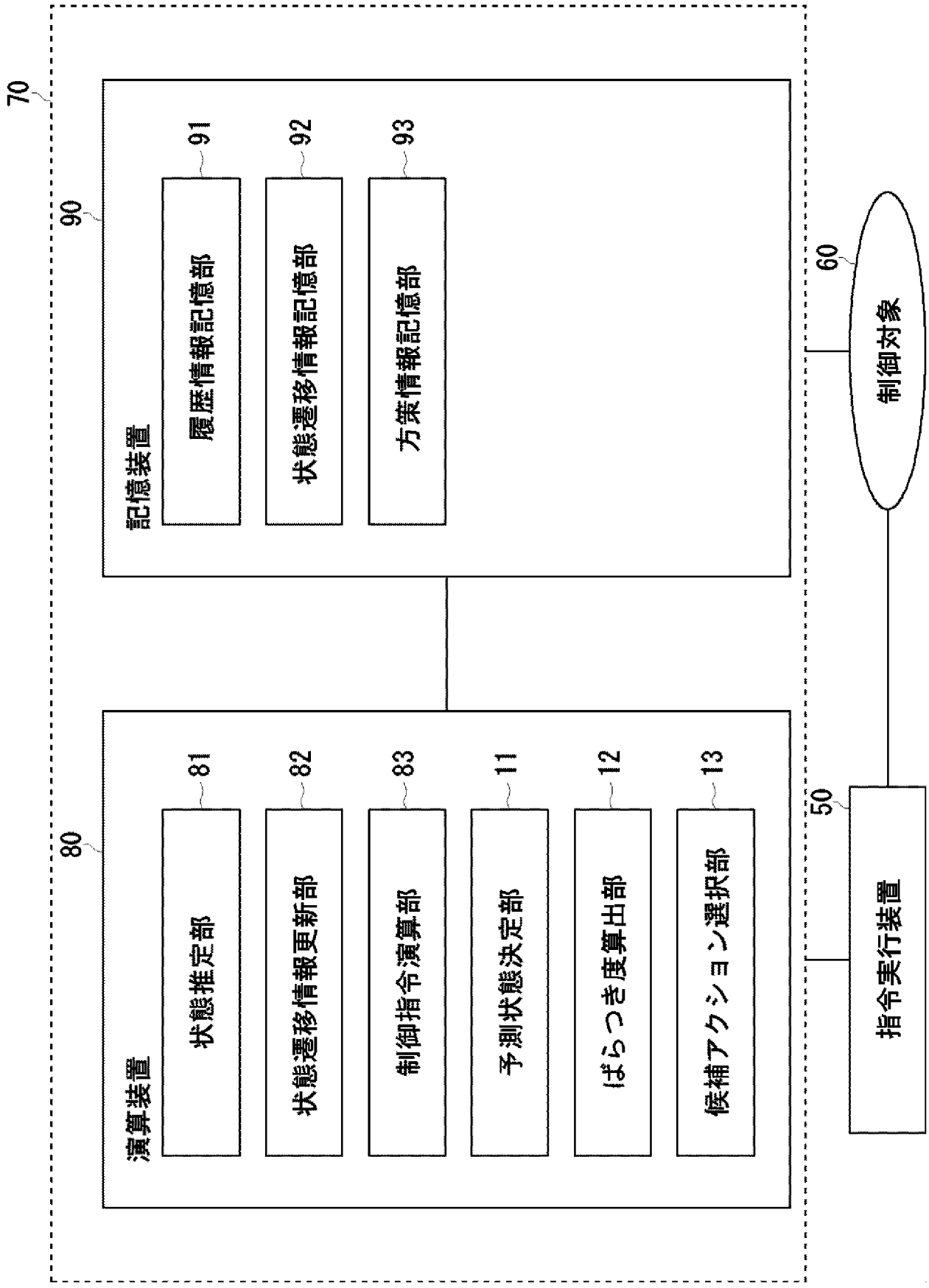
[図2]


20
40
記憶装置
41
履歴情報記憶部
42
状態遷移情報記憶部
43
方策情報記憶部
30
演算装置
31
状態推定部
32
状態遷移情報更新部
33
制御指令演算部
11
予測状態決定部
12
ばらつき度算出部
13
候補アクション選択部
60
制御対象
50
指令実行装置

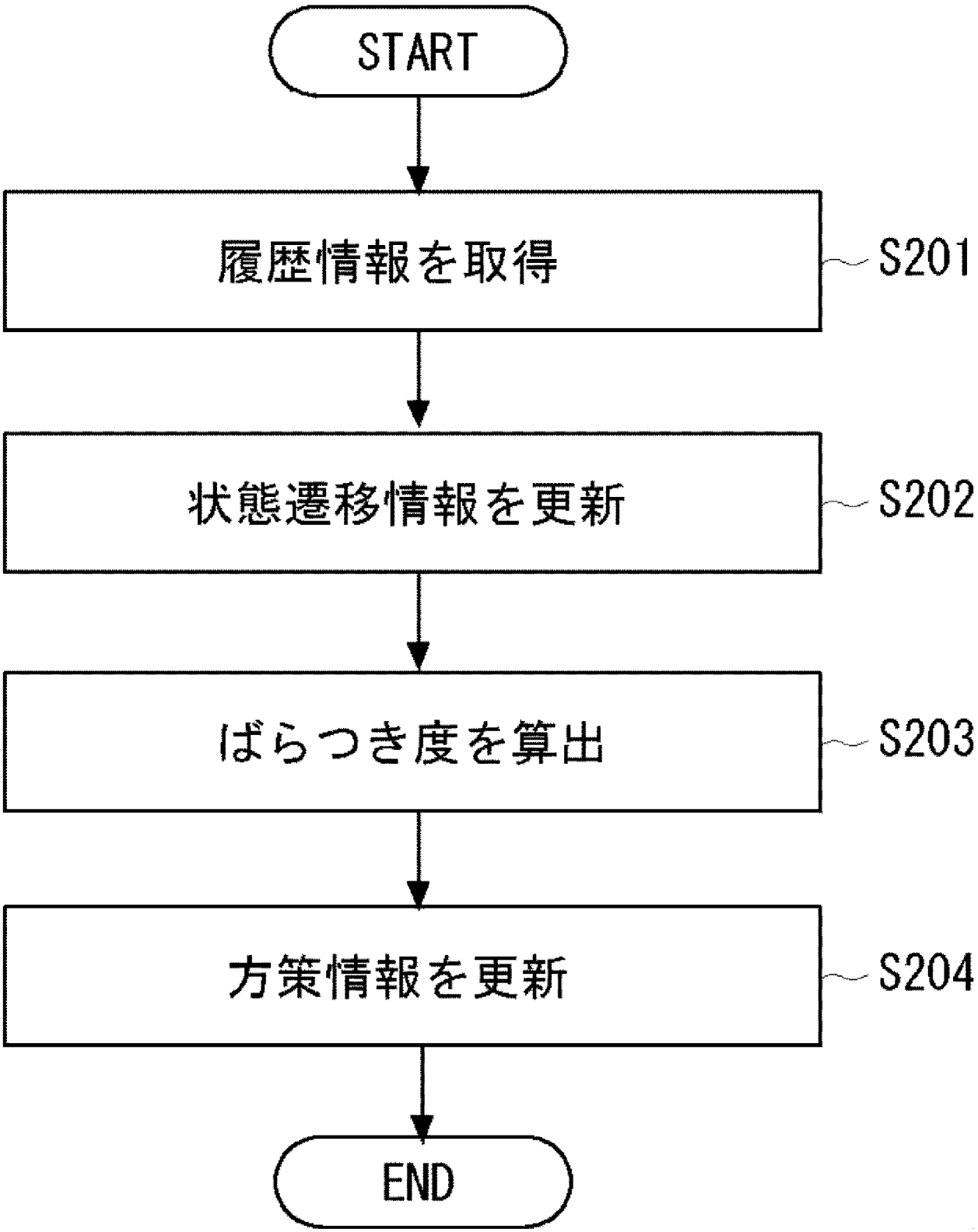
[図3]


START
履歴情報を取得
S101
状態遷移情報を更新
S102
ばらつき度を算出
S103
方策情報を更新
S104
END

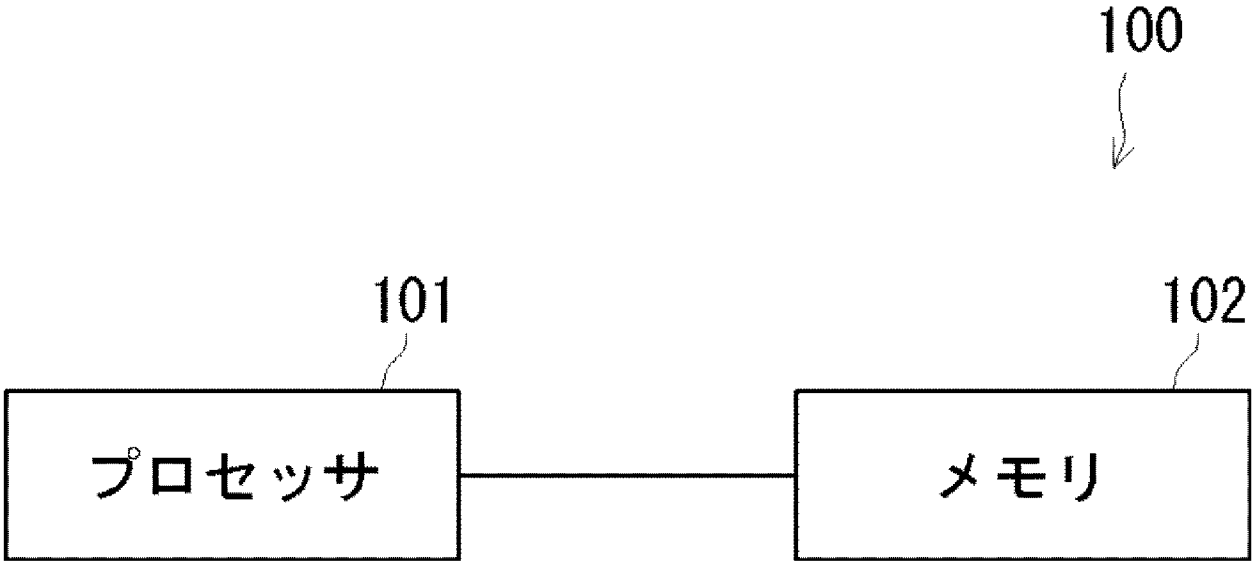
[図4]


70
90
記憶装置
91
履歴情報記憶部
92
状態遷移情報記憶部
93
方策情報記憶部
80
演算装置
81
状態推定部
82
状態遷移情報更新部
83
制御指令演算部
11
予測状態決定部
12
ばらつき度算出部
13
候補アクション選択部
60
制御対象
50
指令実行装置

[図5]


START
履歴情報を取得
S201
状態遷移情報を更新
S202
ばらつき度を算出
S203
方策情報を更新
S204
END

[図6]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2018/045947

A. CLASSIFICATION OF SUBJECT MATTER
Int. Cl. G06N20/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int. Cl. G06N3/00-99/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Published examined utility model applications of Japan 1922-1996
Published unexamined utility model applications of Japan 1971-2019
Registered utility model specifications of Japan 1996-2019
Published registered utility model applications of Japan 1994-2019

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	齋藤雅矩ほか, エージェントの行動履歴の活用によるQ-learningの学習効率向上, 電気学会研究会資料, 07 December 2014, pp. 29-34, in particular, p. 31, right column, line 9 to p. 32, left column, line 41 (SAITO, Masanori et al. Improving efficiency of Q-learning by using the agent's action history. IEEJ Working Group Materials.)	1-4, 9-10 5-8
Y A	HAARNOJA, Tuomas et al., Reinforcement Learning with Deep Energy-Based Policies, arXiv [online], 21 July 2017, [retrieved on 28 January 2019], Retrieved from the Internet: <URL: https://arxiv.org/pdf/1702.08165v2>, in particular, page 2, right column, line 19 to page 3, left column, line 27	1-4, 9-10 5-8

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28.01.2019	12.02.2019
Name and mailing address of the ISA/ Japan Patent Office 3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１８／０４５９４７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06N20/00(2019.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06N3/00－99/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１９年
日本国実用新案登録公報	１９９６－２０１９年
日本国登録実用新案公報	１９９４－２０１９年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Y A	齋藤雅矩ほか，エージェントの行動履歴の活用によるQ-learningの学習効率向上，電気学会研究会資料，2014.12.07，pp.29－34，特に第31頁右欄第９行～第32頁左欄第41行	1－4，9－10 5－8
Y A	HAARNOJA, Tuomas et al., Reinforcement Learning with Deep Energy-Based Policies, arXiv [online], 2017.07.21, [retrieved on 2019.01.28], Retrieved from the Internet: <URL: https://arxiv.org/pdf/1702.08165v2>, 特に第２頁右欄第19行～第３頁左欄第27行	1－4，9－10 5－8

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
２８．０１．２０１９	１２．０２．２０１９

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｂ　３８５６
北元　健太
電話番号　０３－３５８１－１１０１　内線　３５４５