

(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11月26日(26.11.2020)



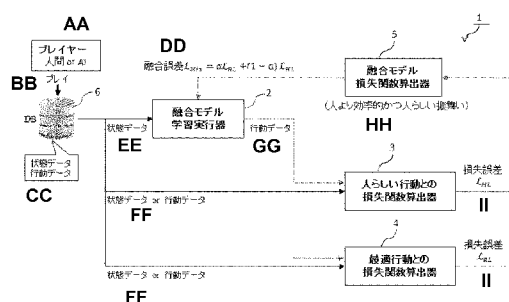
(10) 国際公開番号

WO 2020/235693 A1

- (51) 国際特許分類:
G06N 20/00 (2019.01)
- (21) 国際出願番号: PCT/JP2020/020624
- (22) 国際出願日: 2020年5月25日(25.05.2020)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2019-097222 2019年5月23日(23.05.2019) JP
- (71) 出願人: 国立大学法人神戸大学(NATIONAL UNIVERSITY CORPORATION KOBE UNIVERSITY) [JP/JP]; 〒6578501 兵庫県神戸市灘区六甲台町1番1号 Hyogo (JP). 株式会社エクオス・リサーチ(EQUOS RESEARCH CO., LTD.) [JP/JP]; 〒1010021 東京都千代田区外神田1丁目18番13号 秋葉原ダイビル7F Tokyo (JP).
- (72) 発明者: 松原 崇 (MATSUBARA Takashi); 〒6578501 兵庫県神戸市灘区六甲台町1番1号 国立大学法人神戸大学内 Hyogo (JP). 上原 邦昭(UEHARA Kuniaki); 〒6578501 兵庫県神戸市灘区六甲台町1番1号 国立大学法人神戸大学内 Hyogo (JP). 野本 洋一(NOMOTO Hirokazu); 〒1010021 東京都千代田区外神田1丁目18番13号 秋葉原ダイビル7F 株式会社エクオス・リサーチ内 Tokyo (JP).
- (74) 代理人: 特許業務法人グローバル知財 (THE PATENT CORPORATE BODY GLOBAL INTELLECTUAL PROPERTY); 〒6500021 兵庫県神戸市中央区三宮町3丁目7-6 神戸元町ユニオンビル9F Hyogo (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,

(54) Title: LEARNING METHOD, LEARNING DEVICE, AND LEARNING PROGRAM FOR AI AGENT THAT BEHAVES LIKE HUMAN

(54) 発明の名称: 人らしく振舞うAIエージェントの学習方法、学習装置及び学習プログラム



- 2 Blended model learning executor
3 Calculator of loss function with human-like action
4 Calculator of loss function with optimum action
5 Calculator of blended model loss function
AA Player Human or AI
BB Play
CC State data, action data
DD Blending error
EE State data
FF State data or action data
GG Action data
HH (More efficient and more human-like behavior than human)
II Loss error

(57) Abstract: Provided is a learning method for an agent that combines an optimum action exceeding human capability and a human-like behavior which are acquired by reinforcement learning. Play data by a human expert and state data S_R and action data A_R in the recording of play of the agent of an imitation learning or a reinforcement learning model are inputted. The state data S_R is inputted to the learning executor of a blended model of the reinforcement and imitation learning models and the action data A_F of the blended model is outputted. A first loss error between the action data A_R of the human expert or the agent of the imitation learning model and the action data A_F is calculated. A second loss error between the action data A_{RL} outputted on the basis of the state data S_R by the executor of the reinforcement learning model or an optimum action algorithm and the action data A_F is calculated. A blending error is calculated on the basis of a weight ratio of the first and second loss errors, and the learning executor parameter of the blended model is updated on the basis of the blending error.

DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, KE, KG, KH, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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:

- 国際調査報告 (条約第21条(3))

(57) 要約: 強化学習で獲得される人の能力を超えた最適行動と人らしい振舞いとを両立させたエージェントの学習方法を提供する。人のエキスパートによるプレイデータ、模倣学習又は強化学習モデルのエージェントのプレイの記録における状態データ S_R と行動データ A_R を入力する。強化学習モデルと模倣学習モデルとの融合モデルの学習実行器に対して、状態データ S_R を入力して融合モデルの行動データ A_F を出力させる。人のエキスパート又は模倣学習モデルのエージェントの行動データ A_R と、行動データ A_F との第1の損失誤差を算出する。強化学習モデルの実行器又は最適行動アルゴリズムにより状態データ S_R に基づいて出力される行動データ A_{RL} と、行動データ A_F との第2の損失誤差を算出する。第1及び第2の損失誤差の重み比率に基づいて融合誤差を算出し、融合誤差に基づいて融合モデルの学習実行器のパラメタを更新する。

明 細 書

発明の名称：

人らしく振舞うＡＩエージェントの学習方法、学習装置及び学習プログラム

技術分野

[0001] 本発明は、エージェントＡＩ（人工知能）を訓練する際に、強化学習モデルとの誤差と、人のエキスパートの行動や模倣学習モデルとの誤差とを巧く加味して学習できる方法、装置及びプログラムに関するものである。

背景技術

[0002] 近年、強化学習（Reinforcement learning；以下、“ＲＬ”ということがある）、特に、深層強化学習（ＤＲＬ）は、電子ゲームにおいて利用されるＡＩ（人工知能）、自動車などの車両の自動運転制御、ロボットの自律制御などのアルゴリズムに応用されている。強化学習は、エピソードという単位で学習を繰り返し、環境の中で試行錯誤した結果の報酬を用いて対象となる環境に適応する手法であり方策を最適化することで学習を行うものであり、深層強化学習は、畳み込みニューラルネットワーク（ＣＮＮ）の情報処理力を利用して、画像データなどの高次元入力に基づき、強化学習を行うものである。

[0003] 強化学習や深層強化学習では、環境内におけるエージェントが、現在の状態を観測し、取るべき行動を決定するといった問題が設定され、エージェントは行動を選択することにより環境から報酬を得る。このような環境とエージェントとのやりとりを繰り返しながら、報酬、行動、状態の３つの変数を用いて学習していく。

強化学習や深層強化学習では、エージェントが状態空間と行動空間を持つある環境で行動を取る。タイムステップ毎に、方策が特徴抽出用学習器で状態を特徴ベクトルに変換し、価値関数計算用学習器で特徴ベクトルに対する行動確率を出力する。エージェントが行動を取った後、環境はスカラーの報

酬と次の状態を出力する。最終状態まで行動を取ることを繰り返すと、エピソードが終了する。

[0004] 強化学習や深層強化学習により構築したエージェント A I は、環境と対話しているうちに試行錯誤を重ねて学習して様々な課題が解決でき、人間の能力を遥かに超えることから、ゲーム A I、ロボットの自律制御や自動運転などのアルゴリズムに応用されている（例えば、非特許文献 1～5 を参照。）。

しかしながら、その振る舞いを人間が予測できないことから、例えば人間と共同で作業するロボットや人が乗車する自動運転 A I の実用化に対し、ユーザビリティの点で課題が大きい。また、強化学習や深層強化学習により構築したエージェント A I は、収益を最大化するように訓練されるため高い性能を示すが、実用化する際には、このような性能指標以外のことも考慮する必要がある。例えば、テレビゲームにおいて、ゲームのプレイヤーが操作できないキャラクタである N P C（Non Player Character）を、強化学習のエージェント A I にすると、そのエージェント A I が強すぎるため、プレイヤーがゲームをあまり楽しむことができない可能性がある。また、自動運転に応用する際には、高い性能を目指して訓練された強化学習のエージェント A I は、激しく加減速したり急に曲がったりして、隣接する車や歩行者などに不安を与える恐れがある。そこで、人間らしいエージェント A I を設計する必要がある。

[0005] 一方、模倣学習(Imitation Learning；以下、“I L”ということがある)は、主に人のエキスパートの振舞いを再現する目的で、人間の行動を模倣するエージェント A I を構築するために採用されている。模倣学習においては、エキスパートが従っている方策を最適な方策であると仮定し、エージェント A I の方策がエキスパートの振舞いに近づくように学習が行われる。模倣学習では、人のエキスパートに最適な方策により、状態と行動のペアの系列という形で提供された上で、エージェント A I の方策を観測される状態からエキスパートに取れそうな行動を推測するように訓練されるので、人のエキスパートを模倣し、人らしい振舞いが期待できる。しかしながら、学習され

る方策は提供されたデータに制限されることから、模倣学習のエージェント AI の性能は、人のエキスパートの性能を超えることが殆ど無いといった問題がある（例えば、非特許文献 6～8 を参照。）。

先行技術文献

非特許文献

[0006] 非特許文献1：V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp.529, 2015.

非特許文献2：D. Silver et al., “Mastering the game of go without human knowledge,” Nature, vol. 550, no. 7676, p. 354, 2017.

非特許文献3：B. Vikas, “Deep reinforcement learning approach to autonomous navigation,” 2017.

非特許文献4：T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” CoRR, vol. abs/1509.02971, 2015.

非特許文献5：D. Silver et al., “Deterministic policy gradient algorithms,” in Proceedings of the 31st International Conference on International Conference on Machine Learning, Volume 32, ser. ICML’ 14. JMLR.org, 2014, pp. I-387-I-395.

非特許文献6：S. Ross, G. Gordon, and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 627-635.

非特許文献7：J. Ortega et al., “Imitating human playing styles in super mario bros,” Entertainment Computing, vol. 4, no. 2, pp. 93-104, 2013.

非特許文献8：J. Ho and S. Ermon, “Generative adversarial imitation learning,” CoRR, vol. abs/1606.03476, 2016.

発明の概要

発明が解決しようとする課題

[0007] かかる状況に鑑みて、本発明は、強化学習で獲得される人の能力を超えた効率性の高い最適行動と、人のエキスパートの行動を模倣することで獲得される人らしい振舞いを両立させたエージェントの学習方法、装置及び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

[0008] 上記課題を解決すべく、本発明のエージェントの学習方法は、所定の環境下でエージェントが判断して最適行動をとる振舞いを実現する学習器と人らしい振舞いを実現する学習器とを融合させ、エージェントの行動方策を、人らしく最適行動をとるように最適化する学習方法であって、下記 a) ～ f) のステップを備える。

a) 人のエキスパートによるプレイデータ、又は所定の目的で作成されたエージェントのプレイデータの少なくとも何れかの記録における状態データ S_R と行動データ A_R を入力する入力ステップ。

b) 最適行動をとる振舞いに関わる強化学習モデルと人らしい振舞いに関わる模倣学習モデルとの融合モデルの学習実行器に対して、状態データ S_R を入力して融合モデルの行動データ A_F を出力させる学習ステップ。

c) 行動データ A_R と行動データ A_F との第 1 の損失誤差を算出する第 1 の損失誤差算出ステップ。

d) 強化学習モデルの実行器又は最適行動アルゴリズムにより状態データ S_R に基づいて出力される行動データ A_{RL} と、行動データ A_F との第 2 の損失誤差を算出する第 2 の損失誤差算出ステップ。

e) 第 1 及び第 2 の損失誤差の重み比率に基づいて融合誤差を算出する融合誤差算出ステップ。

f) 融合誤差に基づいて融合モデルの学習実行器のパラメタを更新する更新ステップ。

[0009] 本発明のエージェントの学習方法によれば、強化学習モデルとの誤差と、人のエキスパートの行動や模倣学習モデルとの誤差をフィードバックして、

人らしい振舞いでかつ最適行動をとるエージェント A I を構築することができる。

ここで、強化学習モデルの実行器は、一般的な強化学習と同様に、環境内におけるエージェントが現在の状態を観測し、取るべき行動を決定する方策が設定され、エージェントは価値関数の値が大きくなるように、方策から行動を選択して報酬を得ることを繰り返し、報酬と行動と状態の3つの変数を用いてフィードバックしながら学習して得た実行器である。

[0010] 環境とは、例えば、電子ゲームの画面の画像データの集合や、ロボットや車両に搭載されたカメラの撮影動画の画像データの時系列などの集合である。状態データとは、画像データの集合である環境を観測して得られる個々のゲーム画像、自車両の周辺環境映像、車線位置などのデータである。また、行動データとは、例えば、電子ゲームであればゲームの入力デバイス操作、コントローラの入力（移動方向、ジャンプなど）であり、自動車の自動運転であればハンドル、アクセル、ブレーキ操作などである。

このように、強化学習は、報酬と行動と状態の3つの変数を用いてフィードバックしながら学習する仕組みによって機能するのに対し、模倣学習は、教師エージェントと生徒エージェントを用いた仕組みで機能する。

[0011] また、融合モデルの出力の行動データ A_F は、行動の確率分布で表現される。例えば、コントローラの各入力ボタン操作の確率分布、ハンドル、アクセル、ブレーキ操作の確率分布などである。ここで、確率分布は、最適なものだけの操作出力の場合も、 δ 分布として含まれるとしている。より詳しくは、離散行動の場合は確率分布で表現され、連続行動の場合は最適なものだけが一つだけ出力されることになる。

また、損失誤差とは、例えば損失関数を用いて算出する現時点の出力と期待する出力との誤差をいう。

[0012] 本発明のエージェントの学習方法の入力ステップにおいて、状態データ S_R と行動データ A_R は、下記 1) ～ 5) の場合が存在し、それらの場合に応じて、第 1 及び第 2 の損失誤差算出ステップの処理が異なる。

1) 人のエキスパートによるプレイデータの記録である場合

第1の損失誤差算出ステップは、人のエキスパートによるプレイデータにおける行動データ A_R と、行動データ A_F との誤差を、損失関数を用いて算出する。

第2の損失誤差算出ステップは、強化学習モデルに状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_F との誤差を、損失関数を用いて算出する。

[0013] 2) 模倣学習モデルのエージェントのプレイデータの記録である場合

第1の損失誤差算出ステップは、模倣学習モデルのエージェントのプレイデータにおける行動データ A_R と行動データ A_F との誤差、又は、模倣学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_F との誤差を、損失関数を用いて算出する。

第2の損失誤差算出ステップは、強化学習モデルに状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_F との誤差を、損失関数を用いて算出する。

[0014] 3) 強化学習モデルのエージェントのプレイデータの記録である場合

第1の損失誤差算出ステップは、模倣学習モデルに状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_F との誤差を、損失関数を用いて算出する。

第2の損失誤差算出ステップは、強化学習モデルに状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_F との誤差を、損失関数を用いて算出する。

[0015] 4) 人のエキスパートによるプレイデータの記録における状態データ S_{HE} と行動データ A_{HE} 、及び、強化学習モデルのエージェントのプレイデータの記録における状態データ S_{RL} と行動データ A_{RL} である場合

学習ステップは、状態データ S_{HE} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力させる。

第1の損失誤差算出ステップは、人のエキスパートによるプレイデータにおける行動データ A_{HE} と、行動データ A_{F1} との誤差を、損失関数を用いて算出する。

第2の損失誤差算出ステップは、強化学習モデルに状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_{F2} との誤差を、損失関数を用いて算出する。

[0016] 5) 模倣学習モデルのエージェントのプレイデータの記録における状態データ S_{IL} と行動データ A_{IL} 、及び、強化学習モデルのエージェントのプレイデータの記録における状態データ S_{RL} と行動データ A_{RL} である場合

学習ステップは、状態データ S_{IL} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力させる。

第1の損失誤差算出ステップは、模倣学習モデルのエージェントのプレイデータにおける行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_{F1} との誤差を、損失関数を用いて算出する。

第2の損失誤差算出ステップは、強化学習モデルに状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、行動データ A_{F2} との誤差を、損失関数を用いて算出する。

[0017] 本発明のエージェントの学習方法における知識蒸留によるソフトターゲット化は、エージェントの行動方策の出力が、熱度 T のパラメタを有し、ハードターゲット化とソフトターゲット化に対応するものである。熱度 T が “0” の場合は、行動方策の出力は一つの行動のみが0より大きな確率を持つハードターゲットとなり、行動は一意に決定される。一方、熱度 T が “0” でない場合は、複数の行動が0より大きな確率をもち、行動は確率的に決定される。

[0018] 次に、本発明のエージェントの学習装置について説明する。

本発明のエージェントの学習装置は、所定の環境下でエージェントが判断して最適行動をとる振舞いを実現する学習器と人らしい振舞いを実現する学習器とを融合させ、エージェントの行動方策を人らしく振舞うように最適化

する学習装置であって、下記A)～F)を備える。

A) 人のエキスパートによるプレイデータ、又は所定の目的で作成されたエージェントのプレイデータの少なくとも何れかの記録における状態データ S_R と行動データ A_R を入力する入力部。

B) 最適行動をとる振舞いに関わる強化学習モデルと人らしい振舞いに関わる模倣学習モデルとの融合モデルに対して、状態データ S_R を入力して融合モデルの行動データ A_F を出力する学習実行器。

C) 行動データ A_R と行動データ A_F との第1の損失誤差を算出する第1の損失誤差算出器。

D) 強化学習モデルの実行器又は最適行動アルゴリズムにより状態データ S_R に基づいて出力される行動データ A_{RL} と、行動データ A_F との第2の損失誤差を算出する第2の損失誤差算出器。

E) 第1及び第2の損失誤差の重み比率に基づいて融合誤差を算出する融合誤差算出器。

F) 融合誤差に基づいて前記融合モデルの学習実行器のパラメタを更新する更新部。

[0019] 本発明のエージェントの学習装置の入力部において、上述した本発明のエージェントの学習方法の入力ステップと同様に、上述の1)～5)の場合の状態データ S_R と行動データ A_R に応じて、第1及び第2の損失誤差算出器の算出処理が異なる。

また、本発明のエージェントの学習装置における知識蒸留によるソフトターゲット化は、上述した本発明のエージェントの学習方法と同様に、エージェントの行動方策の出力が、熱度 T のパラメタを有し、ハードターゲット化とソフトターゲット化に対応し得るものであり、熱度 T が“0”の場合は、行動方策の出力は一つの行動のみが0より大きな確率を持つハードターゲットとなり、行動は一意に決定され、一方、熱度 T が“0”でない場合は、複数の行動が0より大きな確率をもち、行動は確率的に決定される。

[0020] 本発明のエージェントの学習プログラムは、上述の本発明のエージェント

の学習方法における全てのステップを、コンピュータに実行させるためのプログラムである。

また本発明のエージェントの学習プログラムは、上述の本発明のエージェントの学習装置における入力部、学習実行器、第1の損失誤差算出器、第2の損失誤差算出器、融合誤差算出器及び更新部として、コンピュータを機能させるためのプログラムである。

発明の効果

[0021] 本発明によれば、強化学習で獲得される人間の能力を超えた効率性の高い最適行動と、人のエキスパートの行動を模倣することで獲得される人らしい振舞いを両立させたエージェントAIを構築することができる。

図面の簡単な説明

- [0022] [図1]本発明のエージェントの学習装置の機能ブロック図
[図2]実施例1のエージェントの学習装置の機能ブロック図
[図3]実施例1のエージェントの学習方法の概略フロー図
[図4]実施例1の融合モデルの実行処理フロー図
[図5]実施例1の人らしい行動との損失誤差関数 L_{HE} の算出処理フロー図
[図6]実施例1の最適行動との損失誤差関数 L_{RL} の算出処理フロー図
[図7]実施例1の融合損失誤差関数 L_{Mix} の算出処理フロー図
[図8]実施例1の融合モデルの学習処理フロー図
[図9]実施例2のエージェントの学習装置の機能ブロック図
[図10]実施例2の人らしい行動との損失誤差関数 L_{IL} の算出処理フロー図
[図11]実施例3のエージェントの学習装置の機能ブロック図
[図12]実施例4のエージェントの学習装置の機能ブロック図
[図13]実施例5のエージェントの学習装置の機能ブロック図
[図14]実施例5のエージェントの学習方法の概略フロー図
[図15]実施例5の融合モデルの実行処理フロー図1
[図16]実施例5の融合モデルの実行処理フロー図2
[図17]実施例5の人らしい行動との損失誤差関数 L_{HE} の算出処理フロー図

[図18]実施例5の最適行動との損失誤差関数 L_{RL} の算出処理フロー図

[図19]実施例6のエージェントの学習装置の機能ブロック図

[図20]実施例3の損失誤差関数 L_{LL} の算出処理フロー図

[図21]ソフトターゲット化処理器の説明図

発明を実施するための最良の形態

[0023] 以下、本発明の実施形態の一例を、図面を参照しながら詳細に説明していく。なお、本発明の範囲は、以下の実施例や図示例に限定されるものではなく、幾多の変更及び変形が可能である。

[0024] 図1は、本発明のエージェントの学習装置の機能ブロック図を示している。学習装置1は、融合モデル学習実行器2、人らしい行動との損失関数算出器3、最適行動との損失関数算出器4、融合モデル損失関数算出器5、及び、データベース(DB)6から構成される。融合モデル学習実行器2は、データベース6から状態データを取得し、融合モデルの実行処理を行い、得られた行動データを、損失関数算出器(3, 4)へ出力する。損失関数算出器(3, 4)は、データベース6から状態データ又は行動データを入力し、入力したデータと融合モデル学習実行器2から入力した行動データをそれぞれ損失誤差関数算出器に取込み、損失誤差関数 L_{HL} 及び損失誤差関数 L_{RL} をそれぞれ算出する。融合モデル損失関数算出器5は、損失関数算出器(3, 4)が算出した損失誤差 L_{HL} 及び損失誤差 L_{RL} を入力し、人より効率的かつ人らしい振舞いとの損失誤差の融合損失誤差関数 L_{Mix} を算出する。融合モデル学習実行器2は、融合モデル損失関数算出器5から融合損失誤差関数 L_{Mix} を取得し、融合モデルの学習処理を行う。

以下の実施例では、融合モデルのタイプ、学習に用いるプレイデータの種別、学習のコントロールの種別の組合せパターンに分けて、本発明のエージェントの学習装置の機能ブロック図を説明する。組合せパターンを表1に纏める。

[0025]

[表1]

実施例 No	結合モデルのタイプ	プレイデータ	コントロール	入力データ	機能ブロック図	備考
1	連続空間タスクに対する 結合モデル	人(エキスパート)	人(エキスパート)	ハードターゲット	図 2	最適行動との損失関数算出器内で、強化学習モデル実行器の出力をハードターゲットとソフトターゲットに変換して損失関数算出に利用
2		模倣学習AI	模倣学習AI	ハードターゲット		
3		模倣学習AI	模倣学習モデル	ハードターゲット	図 1 1	模倣学習AIのプレイデータの行動データを人らしい行動との損失関数算出に直接利用
4		強化学習AI	強化学習モデル	ソフトターゲット		
5	連続空間タスクに対する 結合モデル	人(エキスパート)	人(エキスパート)	ハードターゲット	図 1 2	実施例 3 とプレイデータが異なる以外は同一
		強化学習AI	強化学習モデル	ソフトターゲット		
		強化学習AI	強化学習モデル	ハードターゲット	図 1 3	強化学習AIのプレイデータの行動データを最適行動との損失関数算出に直接利用
		強化学習AI	強化学習モデル	ソフトターゲット		
6	模倣学習と強化学習の ハイブリッド型の結合モデル	模倣学習AI	模倣学習モデル	ハードターゲット	図 1 9	各々の学習モデル実行器の出力をハードターゲットとソフトターゲットに変換して損失関数算出に利用
		強化学習AI	強化学習モデル	ソフトターゲット		
		強化学習AI	強化学習モデル	ソフトターゲット		

実施例 1

[0026] 図2は、実施例1のエージェントの学習装置の機能ブロック図を示している。図2に示すように、学習装置11は、融合モデル学習実行器2、人らしい行動との損失関数算出器3、最適行動との損失関数算出器4、融合モデル損失関数算出器5及びデータベース61から成る。データベース61には、人のエキスパートによるプレイデータが記憶されている。損失関数算出器3には、第1の損失誤差算出器30が設けられ、損失関数算出器4には、強化学習モデル実行器8、ソフトターゲット化処理器9及び第2の損失誤差算出器40が設けられている。

[0027] 学習装置11を用いたエージェントの学習方法について、図3～8を参照しながら説明する。図3は、実施例1のエージェントの学習方法の概略フロー図を示している。図3に示すように、まず、融合モデル学習実行器2において、融合モデルの実行処理を行う（ステップS11）。次に、損失関数算出器3において、人らしい行動との損失誤差関数 L_{HE} の算出処理を行う（ステップS12）。また、これとは別に、最適行動との損失誤差関数 L_{RL} の算出処理を行う（ステップS13）。人らしい行動との損失誤差関数 L_{HE} 及び最適行動との損失誤差関数 L_{RL} を基に、融合モデル損失関数算出器5を用いて、融合損失誤差関数 L_{Mix} の算出処理を行う（ステップS14）。融合モデル学習実行器2において、算出された融合損失誤差関数 L_{Mix} を取得し、融合モデルの学習処理を行う（ステップS15）。融合損失誤差関数 L_{Mix} が所定値以上の場合には（ステップS16）、再度、融合モデル学習実行器2において、融合モデルの実行処理を行う（ステップS11）。

[0028] 次に、図3で示した各処理について説明する。図4は、実施例1の融合モデルの実行処理フロー図を示している。図4に示すように、融合モデル学習実行器2は、データベース61より状態データ S_R を取得する（ステップS111）。融合モデル π へ状態データ S_R を入力し実行する（ステップS112）。融合モデルの行動データ A_F を記憶する（ステップS113）。

[0029] 図5は、実施例1の人らしい行動との損失誤差関数 L_{HE} の算出処理フロー図を示している。図5に示すように、人らしい行動との損失関数算出器3に

おける第1の損失誤差算出器30は、データベース61より人の行動データ A_R を取得する（ステップS121）。また、損失関数算出器3は、融合モデルの行動データ A_F を取得する（ステップS122）。第1の損失誤差算出器30は、取得した A_R と A_F より、損失誤差関数 L_{HE} を算出する（ステップS123）。

[0030] 図6は、実施例1の最適行動との損失誤差関数 L_{RL} の算出処理フロー図を示している。図6に示すように、最適行動との損失関数算出器4において、強化学習モデル実行器8は、データベース61より状態データ S_R を取得する（ステップS131）。強化学習モデル π_{RL} へ状態データ S_R を入力、実行し、行動データ A_R を出力する（ステップS132）。ソフトターゲット化処理器9へ行動データ A_R を入力、実行し、第2の損失誤差算出器40へ行動データ A_{ST2} を出力する（ステップS133）。また、これとは別に、第2の損失誤差算出器40は、融合モデルの行動データ A_F を取得する（ステップS134）。第2の損失誤差算出器40は、取得した行動データ A_{ST2} と行動データ A_F より損失誤差関数 L_{RL} を算出する（ステップS135）。

[0031] ここで、ソフトターゲット化処理器9について図21を参照して説明する。ソフトターゲット化処理器9における知識蒸留によるソフトターゲット化は、前述のとおり、エージェントの行動方策の出力が、熱度 T のパラメタを有し、ハードターゲット化とソフトターゲット化に対応するものである。熱度 $T=0$ の場合は、行動方策の出力は一つの行動のみが0より大きな確率を持つハードターゲットとなり、行動は一意に決定される。一方、熱度 $T>0$ の場合（ T が0でない場合）は、複数の行動が0より大きな確率をもち、行動は確率的に決定される。

ソフトターゲット化処理器について、最適行動との損失関数算出器に用いられる場合を例にして説明する。図21（1）に示すように、強化学習モデル実行器が、状態データを入力し、実行器の出力データ G をソフトターゲット化処理器に出力し、さらにソフトターゲット化処理器が、出力データ L を損失誤差 L_{RL} 算出器に出力する。損失誤差 L_{RL} 算出器は、融合モデルの出力

の行動データと、ソフトターゲット化処理器の出力データとを入力し損失誤差関数を算出する。

[0032] この場合、図21(3)に示すように、ソフトターゲット化処理器のパターンは大きくわけて3通り存在する。まず、ソフトターゲット化処理器内の処理で、熱度 T を処理し、 $T=0$ 、 $T>0$ の場合に分けてそれぞれデータ m とデータ L につき損失誤差関数を算出するパターン（これを“パターンA”とする）、次に、ソフトターゲット化処理器内の処理で、熱度 T を処理し、 $T>0$ とするデータ L につき損失誤差関数を算出するパターン（これを“パターンB”とする）、そして、ソフトターゲット化処理器内の処理で、熱度 T を処理し、 $T=1$ とするデータ G につき損失誤差関数を算出するパターン（これを“パターンC1”とする）と、 $T=0$ とするデータ G につき損失誤差関数を算出するパターン（これを“パターンC2”とする）が存在する。熱度 T によるソフトターゲット化処理は、具体的には、図21(2)に示すグラフのような処理関数を用いて、指数関数のパラメタとして熱度 T を使用している。上記のパターンとデータ名とデータ形式を表2に纏める。

[0033] [表2]

パターン	データ名	データ形式	ソフトターゲット化処理器内
A	データG	ソフトターゲット	—
	データm	ハードターゲット	$T=0$
	データL	熱度 T によるソフトターゲット	$T>0$
B	データG	ソフトターゲット	—
	データL	熱度 T によるソフトターゲット	$T>0$
C1	データG	ソフトターゲット	$T=1$
C2	データG	ハードターゲット	$T=0$

[0034] 図7は、実施例1の融合損失誤差関数 L_{Mix} の算出処理フロー図を示している。図7に示すように、融合モデル損失関数算出器5は、損失誤差関数 L_{HE} 及び L_{RL} を取得する（ステップS141）。融合モデル損失関数算出器5は、トレードオフ係数 α 、 L_{HE} 及び L_{RL} より融合損失誤差関数 L_{Mix} を算出する（ステップS142）。

[0035] 図8は、実施例1の融合モデルの学習処理フロー図を示している。図8に示すように、融合モデル学習実行器2は、融合損失誤差関数 L_{Mix} を取得する（ステップS151）。融合モデル学習実行器2は、融合損失誤差関数 L_{Mix} が小さくなる様に融合モデルのパラメタを変更する（ステップS152）。

実施例 2

[0036] 図9は、実施例2のエージェントの学習装置の機能ブロック図を示している。図9に示すように、学習装置12は、融合モデル学習実行器2、人らしい行動との損失関数算出器3、最適行動との損失関数算出器4、融合モデル損失関数算出器5及びデータベース62から成る。データベース62には、模倣学習モデル実行器7に設けられた模倣学習モデル π_{IL} によるプレイデータが記憶されている。損失関数算出器3には、第1の損失誤差算出器30が設けられ、損失関数算出器4には、強化学習モデル実行器8、ソフトターゲット化処理器9及び第2の損失誤差算出器40が設けられている。

実施例2のエージェントの学習装置12は、使用するデータベースが、模倣学習モデルによるプレイデータが記憶されたデータベース62である点で、人のエキスパートによるプレイデータが記憶されたデータベース61を使用する実施例1の学習装置11とは異なる。そのため、実施例2のエージェントの学習方法では、人らしい行動との損失誤差関数 L_{IL} の算出処理について、実施例1とは異なっている。

[0037] 図10は、実施例2の人らしい行動との損失誤差関数 L_{IL} の算出処理フロー図を示している。図10に示すように、人らしい行動との損失関数算出器3における第1の損失誤差算出器30は、データベース62より模倣学習モデルの行動データ A_{IL} を取得する（ステップS211）。また、損失関数算出器3は、融合モデルの行動データ A_F を取得する（ステップS212）。第1の損失誤差算出器30は、取得した A_{IL} と A_F より、損失誤差関数 L_{IL} を算出する（ステップS213）。

実施例 3

[0038] 図11は、実施例3のエージェントの学習装置の機能ブロック図を示して

いる。図 1 1 に示すように、学習装置 1 3 は、融合モデル学習実行器 2、人らしい行動との損失関数算出器 3、最適行動との損失関数算出器 4、融合モデル損失関数算出器 5 及びデータベース 6 2 から成る。データベース 6 2 には、模倣学習モデル実行器 7 に設けられた模倣学習モデル π_{IL} によるプレイデータが記憶されている。損失関数算出器 4 には、強化学習モデル実行器 8、ソフトターゲット化処理器 9 及び第 2 の損失誤差算出器 4 0 が設けられている。

実施例 3 のエージェントの学習装置 1 3 では、実施例 2 とは異なり、損失関数算出器 3 には、第 1 の損失誤差算出器 3 0 だけではなく、模倣学習モデル実行器 7 及びソフトターゲット化処理器 9 が設けられている。

[0039] 図 2 0 は、実施例 3 の人らしい行動との損失誤差関数 L_{IL} の算出処理フロー図を示している。図 2 0 に示すように、まず、損失関数算出器 3 は、データベース 6 2 から状態データ S_R を取得する（ステップ S 4 1 1）。模倣学習モデル実行器 7 の模倣学習モデル π_{IL} へ状態データ S_R を入力、実行し、行動データ A_{IL} を出力する（ステップ S 4 1 2）。ソフトターゲット化処理器 9 へ行動データ A_{IL} を入力、実行し、行動データ A_{ST1} を出力する（ステップ S 4 1 3）。また、これとは別に、第 1 の損失誤差算出器 3 0 は、融合モデルの行動データ A_F を取得する（ステップ S 4 1 4）。第 1 の損失誤差算出器 3 0 は、取得した行動データ A_{ST1} と行動データ A_F より損失誤差関数 L_{IL} を算出する（ステップ S 4 1 5）。

実施例 4

[0040] 図 1 2 は、実施例 4 のエージェントの学習装置の機能ブロック図を示している。図 1 2 に示すように、学習装置 1 4 は、融合モデル学習実行器 2、人らしい行動との損失関数算出器 3、最適行動との損失関数算出器 4、融合モデル損失関数算出器 5 及びデータベース 6 3 から成る。データベース 6 2 には、強化学習モデル実行器 8 に設けられた強化学習モデル π_{RL} によるプレイデータが記憶されている。損失関数算出器 3 に模倣学習モデル実行器 7、ソフトターゲット化処理器 9 及び第 1 の損失誤差算出器 3 0 が設けられている。

点は、実施例 3 と同様である。また、損失関数算出器 4 には、強化学習モデル実行器 8、ソフトターゲット化処理器 9 及び第 2 の損失誤差算出器 40 が設けられている。

実施例 4 のエージェントの学習装置 14 では、使用するデータベースが、強化学習モデルによるプレイデータが記憶されたデータベース 63 である点で、模倣学習モデルによるプレイデータが記憶されたデータベース 62 を使用する実施例 3 の学習装置 13 と異なるが、その他の点は同様である。

実施例 5

[0041] 図 13 は、実施例 5 のエージェントの学習装置の機能ブロック図を示している。図 13 に示すように、学習装置 15 は、融合モデル学習実行器 2、人らしい行動との損失関数算出器 3、最適行動との損失関数算出器 4、融合モデル損失関数算出器 5 及びデータベース (61, 63) から成る。損失関数算出器 3 には、第 1 の損失誤差算出器 30 が設けられ、損失関数算出器 4 には、第 2 の損失誤差算出器 40 が設けられている。

実施例 5 の学習装置 15 は、人のエキスパートによるプレイデータが記憶されたデータベース 61 と、強化学習モデル実行器 8 に設けられた強化学習モデル π_{RL} によるプレイデータが記憶されたデータベース 63 という 2 種類のデータベースを使用する点で、実施例 1 ~ 4 とは異なっている。

[0042] そこで、学習装置 15 を用いたエージェントの学習方法について、図 14 ~ 18 を参照しながら説明する。図 14 は、実施例 5 のエージェントの学習方法の概略フロー図を示している。図 14 に示すように、まず、融合モデル学習実行器 2 は、融合モデルの実行処理 1 を行う (ステップ S31)。またこれとは別に、融合モデル学習実行器 2 は、融合モデルの実行処理 2 を行う (ステップ S32)。ステップ S31 の融合モデルの実行処理 1 の後に、損失関数算出器 3 において、人らしい行動との損失誤差関数 L_{HE} の算出処理を行う (ステップ S33)。また、ステップ S32 の融合モデルの実行処理 2 の後に、最適行動との損失誤差関数 L_{RL} の算出処理を行う (ステップ S34)。なお、ステップ S33 はステップ S31 の後に行われ、ステップ S34

はステップS 3 2の後に行われればよく、例えば、ステップS 3 1の前にステップS 3 2を行ってもよい。

人らしい行動との損失誤差関数 L_{HE} 及び最適行動との損失誤差関数 L_{RL} を基に、融合モデル損失関数算出器5を用いて、融合損失誤差関数 L_{Mix} の算出処理を行う（ステップS 3 5）。融合モデル学習実行器2において、算出された融合損失誤差関数 L_{Mix} を取得し、融合モデルの学習処理を行う（ステップS 3 6）。融合損失誤差関数 L_{Mix} が所定値以上の場合には（ステップS 3 7）、再度、融合モデル学習実行器2において、融合モデルの実行処理1及び2を行う（ステップS 3 1、ステップS 3 2）。

[0043] 次に、図1 4で示した各処理について説明する。図1 5は、実施例5の融合モデルの実行処理フロー図1を示している。図1 5に示すように、融合モデル学習実行器2は、データベース6 1（DB 1）より状態データ S_{R1} を取得する（ステップS 3 1 1）。融合モデル π へ状態データ S_{R1} を入力し実行する（ステップS 3 1 2）。融合モデルの行動データ A_{F1} を記憶する（ステップS 3 1 3）。

[0044] 図1 6は、実施例5の融合モデルの実行処理フロー図2を示している。図1 6に示すように、融合モデル学習実行器2は、データベース6 3（DB 2）より状態データ S_{R2} を取得する（ステップS 3 2 1）。融合モデル π へ状態データ S_{R2} を入力し実行する（ステップS 3 2 2）。融合モデルの行動データ A_{F2} を記憶する（ステップS 3 2 3）。

[0045] 図1 7は、実施例5の人らしい行動との損失誤差関数 L_{HE} の算出処理フロー図を示している。図1 7に示すように、人らしい行動との損失関数算出器3における第1の損失誤差算出器3 0は、データベース6 1より人の行動データ A_R を取得する（ステップS 3 3 1）。また、損失関数算出器3は、融合モデルの行動データ A_{F1} を取得する（ステップS 3 3 2）。第1の損失誤差算出器3 0は、取得した A_R と A_{F1} より、損失誤差関数 L_{HE} を算出する（ステップS 3 3 3）。

[0046] 図1 8は、実施例5の最適行動との損失誤差関数 L_{RL} の算出処理フロー図

を示している。図18に示すように、最適行動との損失関数算出器4は、データベース63（DB2）より強化学習モデルの行動データ A_{RL} を取得する（ステップS341）。また、これとは別に、第2の損失誤差算出器40は、融合モデルの行動データ A_{F2} を取得する（ステップS342）。第2の損失誤差算出器40は、取得した行動データ A_{RL} と行動データ A_{F2} より損失誤差関数 L_{RL} を算出する（ステップS343）。

[0047] なお、図14に示すステップS35の融合損失誤差関数 L_{Mix} の算出処理については、図7に示す実施例1の融合損失誤差関数 L_{Mix} の算出処理フロー図と同様の処理を行う。また、図14に示すステップS36の融合モデルの学習処理については、図8に示す実施例1の融合モデルの学習処理フロー図と同様の処理を行う。

実施例 6

[0048] 図19は、実施例6のエージェントの学習装置の機能ブロック図を示している。図19に示すように、学習装置16は、融合モデル学習実行器2、人らしい行動との損失関数算出器3、最適行動との損失関数算出器4、融合モデル損失関数算出器5及びデータベース（62，63）から成る。損失関数算出器3には、模倣学習モデル実行器7、ソフトターゲット化処理器9及び第1の損失誤差算出器30が設けられ、損失関数算出器4には、強化学習モデル実行器8、ソフトターゲット化処理器9及び第2の損失誤差算出器40が設けられている。

実施例6の学習装置16は、模倣学習モデル実行器7に設けられた模倣学習モデル π_{IL} によるプレイデータが記憶されたデータベース62（DB1）と、強化学習モデル実行器8に設けられた強化学習モデル π_{RL} によるプレイデータが記憶されたデータベース63（DB2）という2種類のデータベースを使用する点で、データベース（61，63）を使用する実施例5とは異なっている。

実施例 7

[0049] （融合モデルの性能評価について）

上述の実施例で説明したとおり、強化学習モデルの高い性能を保ったまま人らしいエージェント A I を学習するという学習方法は、人の性能を凌駕する効率的な最適行動をとるエージェント A I を学習する処理と、人らしく行動を選択するエージェント A I を学習する処理の 2 つから構成されている。

各処理は、それぞれの強化学習と模倣学習の課題として取り組まれている。そこで、本発明では、強化学習と模倣学習の融合モデルについて、離散行動空間の場合の方策の蒸留に基づき、連続行動空間の場合は敵対模倣学習に基づく方法となっている。

また、 π^* を強化学習モデルによる最適な方策、 π_{HE} を人（エキスパート）の方策とし、これらの 2 つの方策の比率を決めるパラメタを $\alpha \in (0, 1)$ とすると、目的関数は下記の式ようになる。

[0050] [数1]

$$\mathcal{L}_{mix}(\pi) = \alpha \mathcal{L}_{RL}(\pi) + (1 - \alpha) \mathcal{L}_{HE}(\pi)$$

[0051] 離散行動空間の場合、模倣学習の目的関数を、模倣学習の先行する研究に従って、以下の交差エントロピー損失として定義する。

[0052] [数2]

$$\mathcal{L}_{\pi^*}(\pi) = \mathbb{E}_s \left[- \sum_a \pi^*(a|s) \log \pi(a|s) \right]$$

[0053] 人（エキスパート）の方策 π_{HE} は、数理モデルとして定義するのが難しく、実験的にサンプリングされたデータの上で学習を行うことにする。 π_{HE} からソフトターゲットが得られないにも関わらず、方策の蒸留には、ハードターゲットとソフトターゲットの重付き平均を計算することでより良い性能が得られる。したがって、人（エキスパート）に提供されるデータをハードターゲットとし、学習済みモデルの方策 $\pi^{(T)}_{RL}$ の出力を、熱度 T で調整したものをソフトターゲットとする。最終的に、損失関数は以下ようになる。

[0054]

[数3]

$$\begin{aligned}\mathcal{L}_{mix}(\pi) = & \alpha \mathbb{E}_{\pi_{RL}} \left[- \sum_a \pi_{RL}^{(T)}(a|s) \log \pi(a|s) \right] \\ & + (1 - \alpha) \mathbb{E}_{\pi_{HE}} \left[- \sum_a \pi_{HE}(a|s) \log \pi(a|s) \right]\end{aligned}$$

[0055] 連続行動空間の場合には、模倣学習方法として、既知の方法のGAIL (Generative Adversarial Imitation Learning)法を用いる (GAIL法については、非特許文献8を参照。)。GAIL法は教師モデル π からサンプリングされた軌道 $\tau \sim \pi$ を必要とする。GAIL法における識別器 D_w に最大化される目的関数と生徒モデルに最小化される目的関数は以下ようになる。

[0056] [数4]

$$\mathcal{L}_{\pi^*}(\pi) = \mathbb{E}_{\tau \sim \pi} [\log(D_w(s, a))] + \mathbb{E}_{\tau^* \sim \pi^*} [\log(1 - D_w(s, a))]$$

[0057] ここで、 τ は生徒モデルからサンプリングされた軌道 $\tau \sim \pi$ である。融合化モデルとする際には、教師モデルを人 (エキスパート) と強化学習モデルにするため、それぞれのエキスパートから軌道 $\tau_{HE} \sim \pi_{HE}$ と、 $\tau_{RL} \sim \pi_{RL}$ をサンプリングする。さらに、融合の損失関数を以下のように置き換えることができる。直感的には、識別器 D_w は、人 (エキスパート) と強化学習モデルの方策間の融合方策を認めるように学習され、この識別器を騙せるように訓練される生徒モデル π が融合方策に近づき、両方のエキスパートの長所を模倣すると期待される。

[0058] [数5]

$$\begin{aligned}\mathcal{L}_{mix}(\pi) = & \mathbb{E}_{\tau \sim \pi} [\log(D_w(s, a))] \\ & + \alpha \mathbb{E}_{\tau_{RL} \sim \pi_{RL}} [\log(1 - D_w(s, a))] \\ & + (1 - \alpha) \mathbb{E}_{\tau_{HE} \sim \pi_{HE}} [\log(1 - D_w(s, a))]\end{aligned}$$

[0059] 実施した実験について説明する。

(実験1) Atari 2600 ゲーム (Gopher)

実施例1の学習装置を、まず離散行動空間のGopherという名称のAtari 2600システムのゲームに適用した。このゲームは、地下から地上に出てくるねず

み (Gopher) が人参を取れないように、ゲームのユーザが農夫として行動し (振舞い)、左右に動いたり穴を埋めたりすることである。人 (エキスパート) と訓練済みの学習モデルがそれぞれ55000のフレームを提供した上、訓練セットを50000、テストセットを5000として学習を行った。特に、生徒モデルを訓練するために学習率 10^{-4} のAdam optimizerとDropout率0.5を利用した。また、知識蒸留の熱度を $T = 0.1$ 、トレードオフ係数を $\alpha = 0.93$ として、融合モデルの学習を行った。

[0060] (実験2) Torcs

Torcs (Wymann, B., “The open racing car simulator”, (2015)を参照) は、自動運転の研究で最もよく利用されるシミュレータの1つである。このTorcsを用いた実験は、GymTorcs環境をベースにした。エージェントA1の観測空間は、車から端までの距離、敵の車までの距離、現在の速度や加速度など、全体で65の連続値から成っている。行動空間は、2つの要素「左右」と「加減速」から成っており、取りうる値は $[-1.0, 1.0]$ の範囲に限られる。

報酬関数は、走った距離にし、強化学習モデルをOpenAI Baselinesを基に訓練した。さらに、人らしさが見分けられる状況が現れるように、Torcsのシミュレータに停車ボットをした上、人 (エキスパート) に60秒の220エピソードをプレイさせ、そのデータを収集した。単なる模倣学習エージェントを、人 (エキスパート) のデータ上で、強化学習モデルの訓練と同じく、OpenAIのGAIL法を用いて訓練を行った。最後に、GAIL法における学習器の更新を実施し、両方のエキスパートの影響を等しくするためにトレードオフ係数 $\alpha = 0.5$ にして、融合モデルの学習を行った。

(実験3) Appleゲーム

Appleゲームは、ランダムに1つ画面上に現われたリンゴをプレイヤーが収集し、収集されたリンゴの数でスコアを出すものである。

[0061] (3) 人間らしさの感性試験

各モデルの評価以外は、モデルの人らしさを評価するためにダブル・ブラインドで感性試験を実施した。その試験は男性23人、女性3人の計26人

の審査員を対象にした。年齢は27から59歳、平均年齢は44歳である。感性試験の審査員全員、本調査以前に、本発明内容に関する資料には接触がなかった。初めに、審査員に各ゲームのルールを説明し、人らしい行動（振舞い）を理解してもらえるように各ゲームの体験会を実施した。調査の内容は、各審査員、ゲーム毎に2本の動画（Gopherの場合15秒、Torcsの場合30秒）を提供し、人間かAIかの判断とその理由を依頼した。

[0062] 実験結果について説明する。

(1) Atari 2600 ゲーム (Gopher)

性能に関して、点数の高い順に、強化学習モデル（比較例3）、次に実施例の融合モデル、最後に人（比較例1）と模倣学習モデル（比較例2）の順番になった。実施例の融合モデルは、強化学習モデル（比較例3）に提供されたターゲットを $\alpha = 0.8$ で優先したにも関わらず、スコアの向上が3点しかなく、単体の強化学習モデルの点数との差が大きい。感性試験に関して、強化学習モデル（比較例3）は、あまり人間らしくないと判断されたが、実施例の融合モデルは点数だけではなく、人らしさでも、人（比較例1）とその模倣（比較例2）より高いスコアを示した。

融合モデルは、強化学習の目標に向かった学習傾向と人（エキスパート）の振舞いを学習できたことがわかる。なお、意外にも、融合モデルは人（エキスパート）よりも、人らしいと判断された。その理由を解明するために、審査員のコメント分析によってよく現れた感想は、“無駄な動きが少ない”、“動きが細かい、プログラム感が動きにある”、“穴を順番に埋めようとする”であった。従って、人（エキスパート）は、特にゲームをあまりしない審査員に高い性能が期待されていなかったと考えられる。点数の結果を表3に纏める。

[0063]

[表3]

エージェント	点数		調査
	平均	偏差	判断率(%)
実施例（模倣学習＋強化学習）	26.05	24.31	59.62
比較例 1（人）	23.87	19.81	55.70
比較例 2（模倣学習）	23.91	23.79	59.62
比較例 3（強化学習）	40.30	36.81	32.69

[0064] （2）Torcs

性能の評価に関して、まず、それぞれの人（エキスパート）、GAILによる人の模倣学習、強化学習のDDPGと融合モデルの点数を比較した。点数の結果を表4に纏める。実験より、GAILは訓練済みの強化学習モデルや決定論的なボットの模倣に優れているが、人（エキスパート）の模倣の効率は意外に低いことが観測された。それは、人（エキスパート）の方策は複雑で、基本的なニューラルネットワークでの扱いが困難だと推定される。一方で、実施例の融合モデルは、例えば、強化学習モデル（比較例3）の高速度や、人（比較例1）の曲がり方という特徴の模倣に成功した。さらに、人（比較例1）と強化学習モデル（比較例3）のように全体のトラックを走れるようになった。

人らしいと判断された割合が低い順に、まず強化学習モデル（比較例3）は“走るのが速すぎる”や“高速で角を曲がる”という理由で、あまり人間らしくないと判断された。意外にも、人（比較例1）は人らしくないと判断された。同じ動画に対して、審査員のコメントが多様であるが、より低い性能を示した人の模倣エージェントAIが、人（エキスパート）より人らしいと判断されたため、Gopherと同じく、人（比較例1）に示された性能は高いといえる。最後に、融合モデルは最も人らしいと判断された上、強化学習モデル（比較例3）に近い性能を示した。

[0065] [表4]

エージェント	点数($\times 10^3$)		調査
	平均	偏差	判断率(%)
実施例(模倣学習+強化学習)	36.63	1.32	61.54
比較例1(人)	40.17	3.63	50.00
比較例2(模倣学習)	23.99	1.16	51.92
比較例3(強化学習)	40.45	0.43	30.77

[0066] (実験3) Appleゲーム

Appleゲームは単純な原理、すなわち、プレイヤーのアバターを出現したリングの位置に移動させるゲームである。そして、強化学習が最も高いスコアを達成し、続いて、実施例の学習方法、そして最後に、人(比較例1)とその模倣学習が続いた。

人らしさに関する限り、人のエージェント(比較例1)が最もよかった。実施例の学習方法は、スコアにおいて人(比較例1)及びその模倣学習モデルのエージェント(比較例2)を上回りながら、強化学習モデルのエージェント(比較例3)よりも人らしい行動を示し、その後に人(比較例1)が続いた。この結果から、実施例の学習方法は、このゲームにおける人らしい行動と高性能のバランスをとることがわかる。

[0067] [表5]

エージェント	点数($\times 10^3$)		調査
	平均	偏差	判断率(%)
実施例(模倣学習+強化学習)	22.02	3.70	54.0
比較例1(人)	18.71	2.86	64.0
比較例2(模倣学習)	17.57	4.37	44.0
比較例3(強化学習)	36.27	5.44	8.0

産業上の利用可能性

[0068] 本発明は、自動車の自動運転、工業用ロボットアームの自動制御など幅広い分野におけるエージェントの学習に有用である。

符号の説明

- [0069] 1, 11～16 学習装置
- 2 融合モデル学習実行器
 - 3, 4 損失関数算出器
 - 5 融合モデル損失関数算出器
 - 6, 61～63 データベース
 - 7 模倣学習モデル実行器
 - 8 強化学習モデル実行器
 - 9 ソフトターゲット化処理器
 - 30 第1の損失誤差算出器
 - 40 第2の損失誤差算出器

請求の範囲

[請求項1]

所定の環境下でエージェントが判断して最適行動をとる振舞いを実現する学習器と人らしい振舞いを実現する学習器とを融合させ、前記エージェントの行動方策を、人らしく最適行動をとるように最適化する学習方法であって、

人のエキスパートによるプレイデータ、又は所定の目的で作成されたエージェントのプレイデータの少なくとも何れかの記録における状態データ S_R と行動データ A_R を入力する入力ステップと、

前記最適行動をとる振舞いに関わる強化学習モデルと前記人らしい振舞いに関わる模倣学習モデルとの融合モデルの学習実行器に対して、前記状態データ S_R を入力して融合モデルの行動データ A_F を出力させる学習ステップと、

前記行動データ A_R と前記行動データ A_F との第1の損失誤差を算出する第1の損失誤差算出ステップと、

前記強化学習モデルの実行器又は最適行動アルゴリズムにより前記状態データ S_R に基づいて出力される行動データ A_{RL} と、前記行動データ A_F との第2の損失誤差を算出する第2の損失誤差算出ステップと、

第1及び第2の損失誤差の重み比率に基づいて融合誤差を算出する融合誤差算出ステップと、

前記融合誤差に基づいて前記融合モデルの学習実行器のパラメタを更新する更新ステップ、を備えることを特徴とするエージェントの学習方法。

[請求項2]

前記入力ステップにおいて、前記状態データ S_R と行動データ A_R が、人のエキスパートによるプレイデータの記録である場合には、上記の第1の損失誤差算出ステップは、

人のエキスパートによるプレイデータにおける前記行動データ A_R と、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出ステップは、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出することを特徴とする請求項1に記載のエージェントの学習方法。

[請求項3]

前記入力ステップにおいて、前記状態データ S_R と行動データ A_R が、模倣学習モデルのエージェントのプレイデータの記録である場合には、

上記の第1の損失誤差算出ステップは、

前記模倣学習モデルのエージェントのプレイデータにおける前記行動データ A_R と前記行動データ A_F との誤差、又は、前記模倣学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出ステップは、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出することを特徴とする請求項1に記載のエージェントの学習方法。

[請求項4]

前記入力ステップにおいて、前記状態データ S_R と行動データ A_R が、強化学習モデルのエージェントのプレイデータの記録である場合には、

上記の第1の損失誤差算出ステップは、

前記模倣学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出ステップは、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行

動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出することを特徴とする請求項 1 に記載のエージェントの学習方法。

[請求項5]

前記入力ステップにおいて、前記状態データ S_R と行動データ A_R が、人のエキスパートによるプレイデータの記録における状態データ S_{HE} と行動データ A_{HE} 、及び、前記強化学習モデルのエージェントのプレイデータの記録における状態データ S_{RL} と行動データ A_{RL} である場合には、

前記学習ステップは、前記状態データ S_{HE} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力させ、

上記の第 1 の損失誤差算出ステップは、

人のエキスパートによるプレイデータにおける前記行動データ A_{HE} と、前記行動データ A_{F1} との誤差を、損失関数を用いて算出し、

上記の第 2 の損失誤差算出ステップは、

前記強化学習モデルに前記状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_{F2} との誤差を、損失関数を用いて算出することを特徴とする請求項 1 に記載のエージェントの学習方法。

[請求項6]

前記入力ステップにおいて、前記状態データ S_R と行動データ A_R が、前記模倣学習モデルのエージェントのプレイデータの記録における状態データ S_{IL} と行動データ A_{IL} 、及び、前記強化学習モデルのエージェントのプレイデータの記録における状態データ S_{RL} と行動データ A_{RL} である場合には、

前記学習ステップは、前記状態データ S_{IL} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力させ、

上記の第 1 の損失誤差算出ステップは、

前記模倣学習モデルのエージェントのプレイデータにおける前記行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_{F1} との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出ステップは、

前記強化学習モデルに前記状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_{F2} との誤差を、損失関数を用いて算出することを特徴とする請求項1に記載のエージェントの学習方法。

[請求項7] 前記知識蒸留によるソフトターゲット化は、エージェントの行動方策の出力が、熱度 T のパラメタを有し、ハードターゲット化とソフトターゲット化に対応し得ることを特徴とする請求項2～6の何れかに記載のエージェントの学習方法。

[請求項8] 所定の環境下でエージェントが判断して最適行動をとる振舞いを実現する学習器と人らしい振舞いを実現する学習器とを融合させ、前記エージェントの行動方策を人らしく振舞うように最適化する学習装置であって、

人のエキスパートによるプレイデータ、又は所定の目的で作成されたエージェントのプレイデータの少なくとも何れかの記録における状態データ S_R と行動データ A_R を入力する入力部と、

前記最適行動をとる振舞いに関わる強化学習モデルと前記人らしい振舞いに関わる模倣学習モデルとの融合モデルに対して、前記状態データ S_R を入力して融合モデルの行動データ A_F を出力する学習実行器と、

前記行動データ A_R と前記行動データ A_F との第1の損失誤差を算出する第1の損失誤差算出器と、

前記強化学習モデルの実行器又は最適行動アルゴリズムにより前記状態データ S_R に基づいて出力される行動データ A_{RL} と、前記行動データ A_F との第2の損失誤差を算出する第2の損失誤差算出器と、

第1及び第2の損失誤差の重み比率に基づいて融合誤差を算出する融合誤差算出器と、

前記融合誤差に基づいて前記融合モデルの学習実行器のパラメタを更新する更新部、を備えることを特徴とするエージェントの学習装置。

[請求項9] 前記入力部における前記状態データ S_R と行動データ A_R に応じて、下記1)～5)の何れかの算出処理を行う構成を備えることを特徴とする請求項8に記載のエージェントの学習装置：

1) 前記状態データ S_R と行動データ A_R が、人のエキスパートによるプレイデータの記録である場合には、

上記の第1の損失誤差算出器は、

人のエキスパートによるプレイデータにおける前記行動データ A_R と、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出器は、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出し、

2) 前記状態データ S_R と行動データ A_R が、模倣学習モデルのエージェントのプレイデータの記録である場合には、

上記の第1の損失誤差算出器は、

前記模倣学習モデルのエージェントのプレイデータにおける前記行動データ A_R と前記行動データ A_F との誤差、又は、前記模倣学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出器は、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データ

と、前記行動データ A_F との誤差を、損失関数を用いて算出し、

3) 前記状態データ S_R と行動データ A_R が、強化学習モデルのエージェントのプレイデータの記録である場合には、

上記の第1の損失誤差算出器は、

前記模倣学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データ

と、前記行動データ A_F との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出器は、

前記強化学習モデルに前記状態データ S_R を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データ

と、前記行動データ A_F との誤差を、損失関数を用いて算出し、

4) 前記状態データ S_R と行動データ A_R が、人のエキスパートによるプレイデータの記録における状態データ S_{HE} と行動データ A_{HE} 、及び、前記強化学習モデルのエージェントのプレイデータの記録における状態データ S_{RL} と行動データ A_{RL} である場合には、

前記学習実行器は、前記状態データ S_{HE} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力し、

上記の第1の損失誤差算出器は、

人のエキスパートによるプレイデータにおける前記行動データ A_{HE} と、前記行動データ A_{F1} との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出器は、

前記強化学習モデルに前記状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データ

と、前記行動データ A_{F2} との誤差を、損失関数を用いて算出し、

5) 前記状態データ S_R と行動データ A_R が、前記模倣学習モデルのエージェントのプレイデータの記録における状態データ S_{IL} と行動データ A_{IL} 、及び、前記強化学習モデルのエージェントのプレイデ

ータの記録における状態データ S_{RL} と行動データ A_{RL} である場合には、

前記学習実行器は、前記状態データ S_{IL} を入力して行動データ A_{F1} を出力させ、及び、状態データ S_{RL} を入力して行動データ A_{F2} を出力し、

上記の第1の損失誤差算出器は、

前記模倣学習モデルのエージェントのプレイデータにおける前記行動データ A_{IL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_{F1} との誤差を、損失関数を用いて算出し、

上記の第2の損失誤差算出器は、

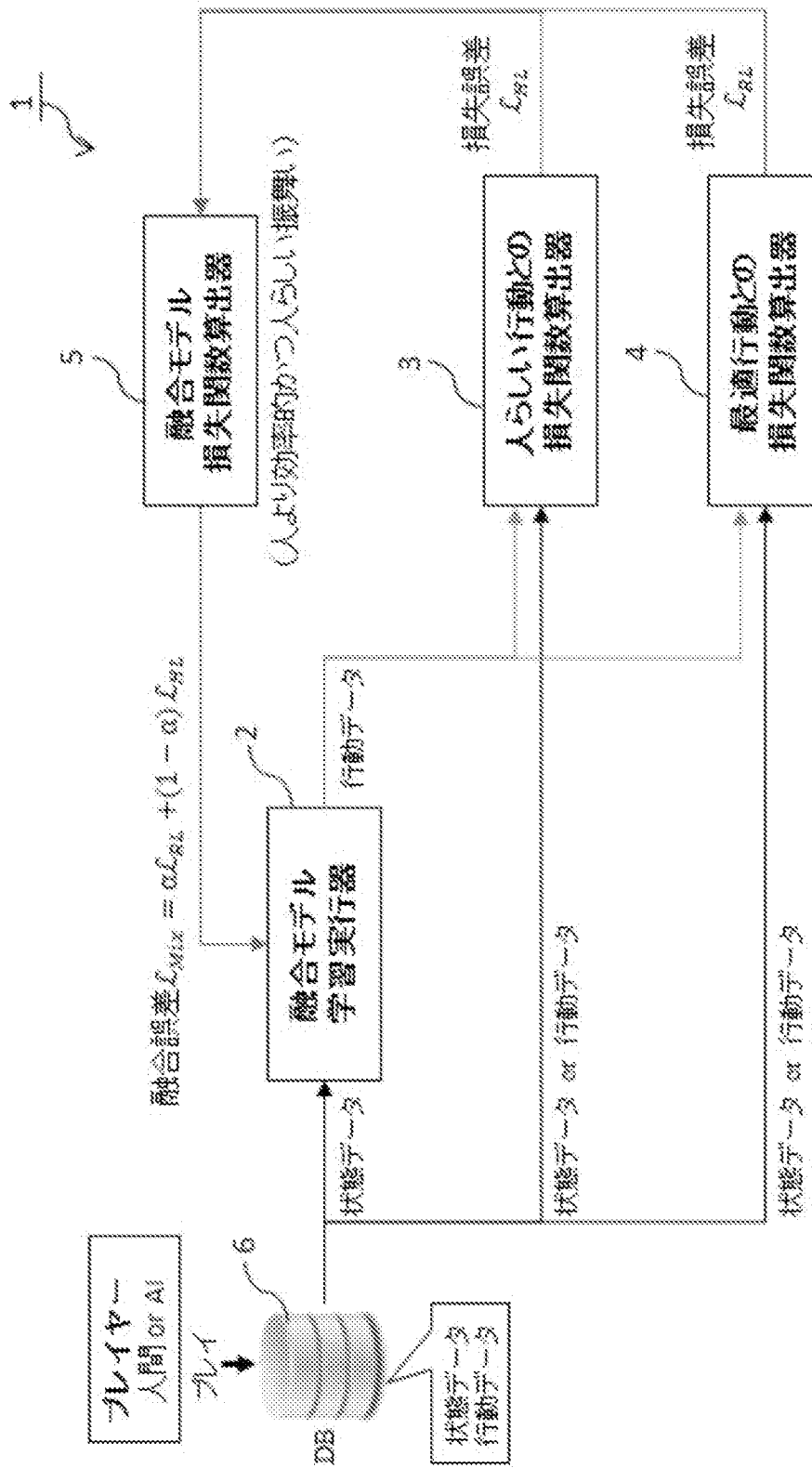
前記強化学習モデルに前記状態データ S_{RL} を入力して出力させた行動データ A_{RL} を知識蒸留によるソフトターゲット化した行動データと、前記行動データ A_{F2} との誤差を、損失関数を用いて算出する。

[請求項10] 前記知識蒸留によるソフトターゲット化は、エージェントの行動方針の出力が、熱度 T のパラメタを有し、ハードターゲット化とソフトターゲット化に対応し得ることを特徴とする請求項9に記載のエージェントの学習装置。

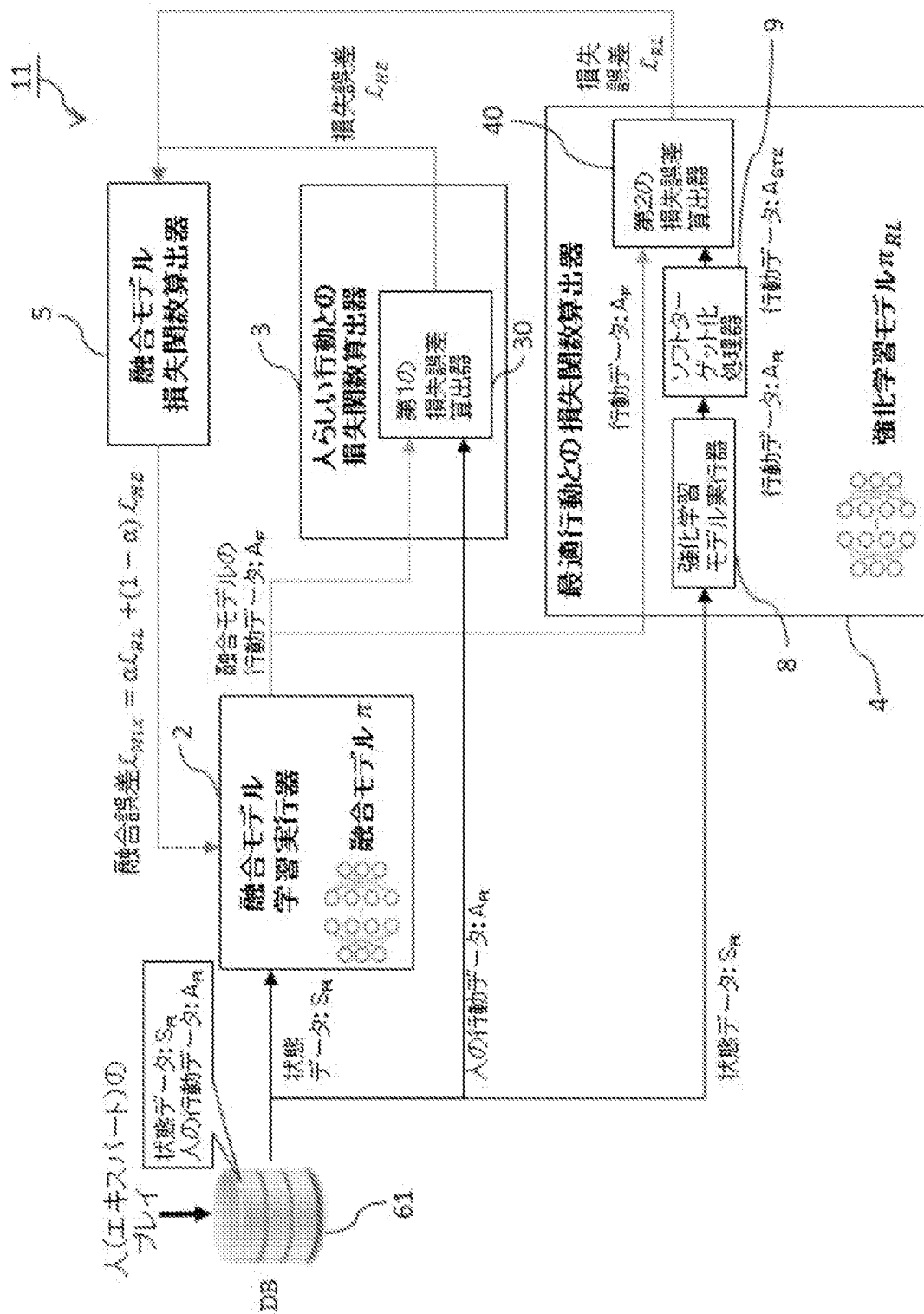
[請求項11] 請求項1～7の何れかのエージェントの学習方法における全てのステップを、コンピュータに実行させるためのエージェントの学習プログラム。

[請求項12] 請求項8～10の何れかのエージェントの学習装置における入力部、学習実行器、第1の損失誤差算出器、第2の損失誤差算出器、融合誤差算出器、及び、更新部として、コンピュータを機能させるためのエージェントの学習プログラム。

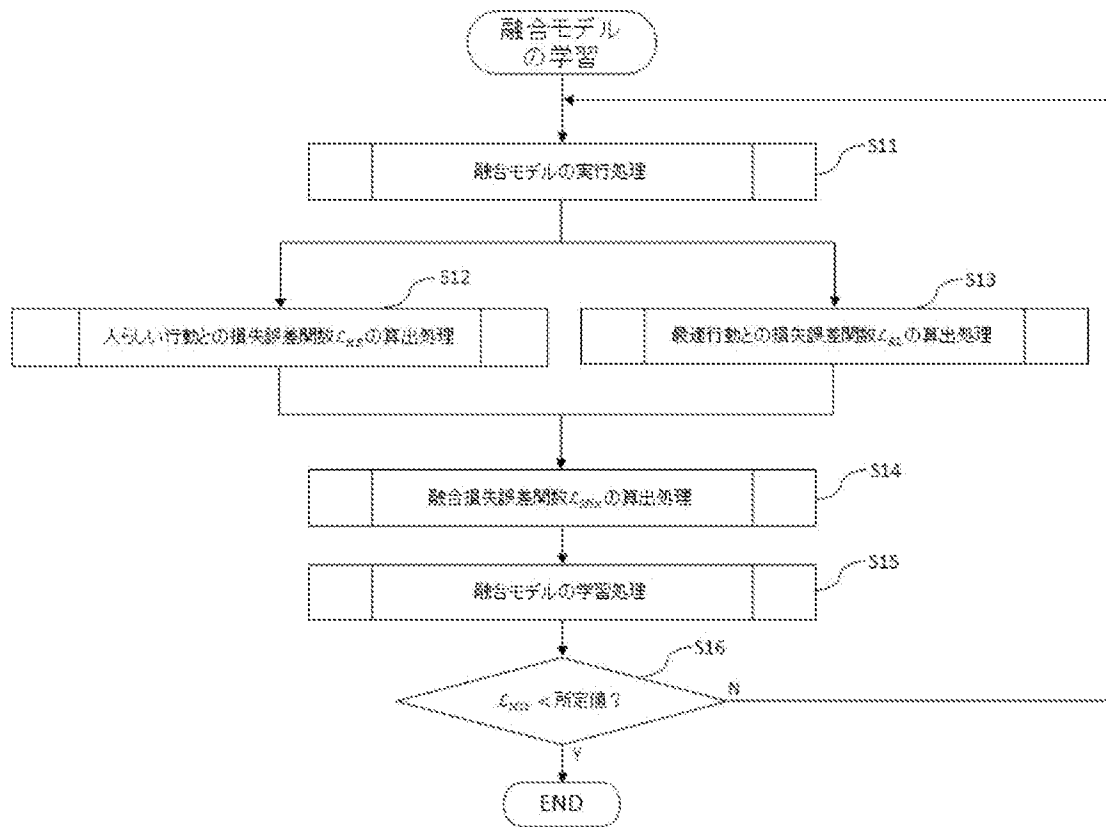
[図1]



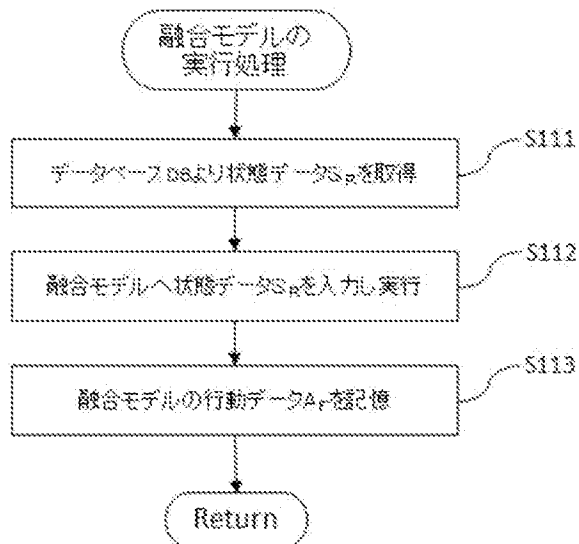
[図2]



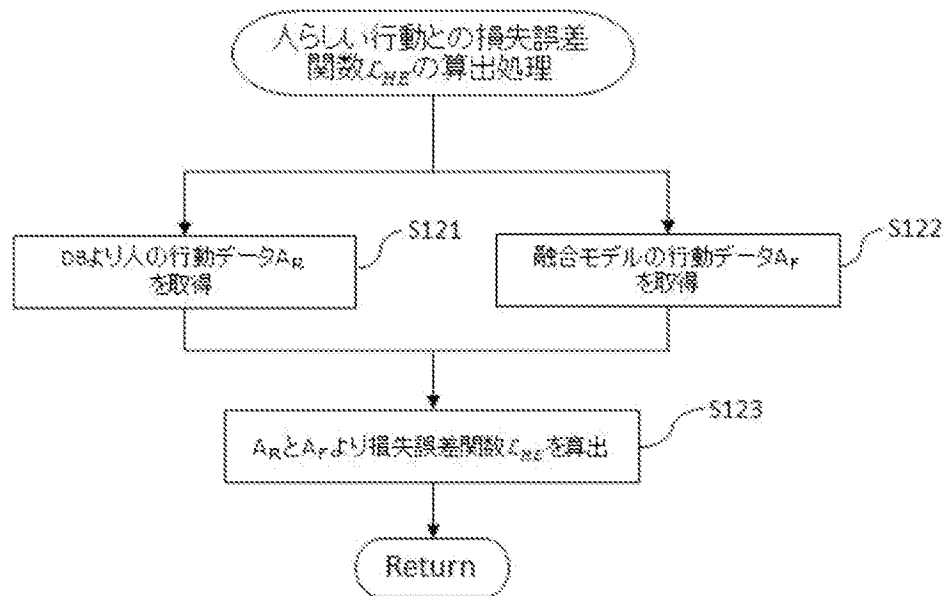
[図3]



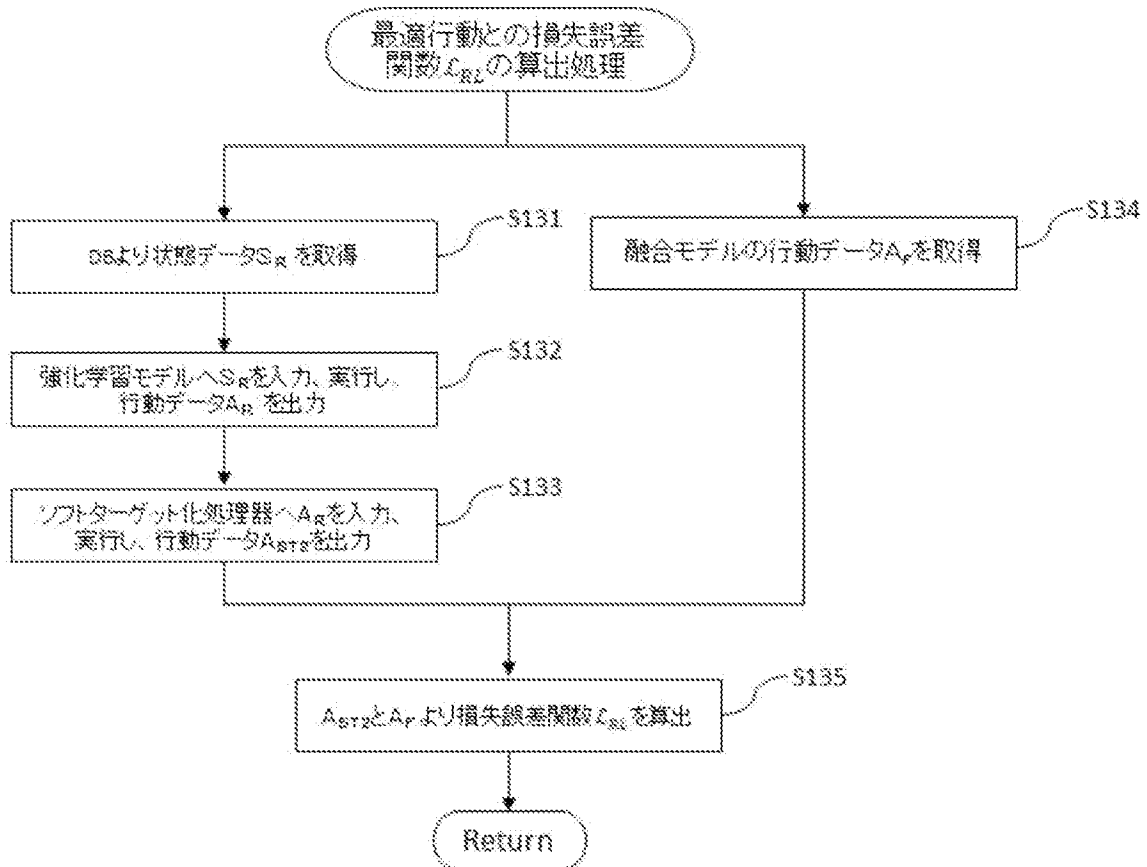
[図4]



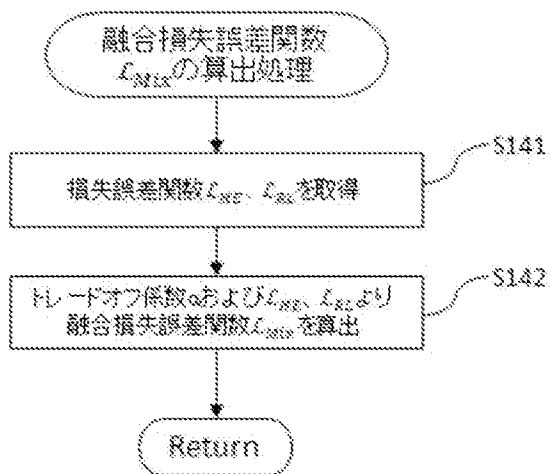
[図5]



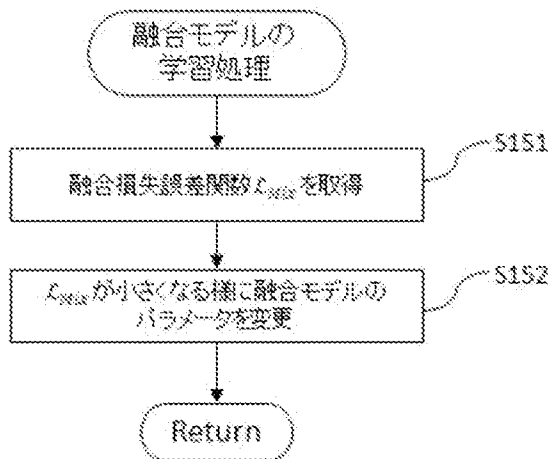
[図6]



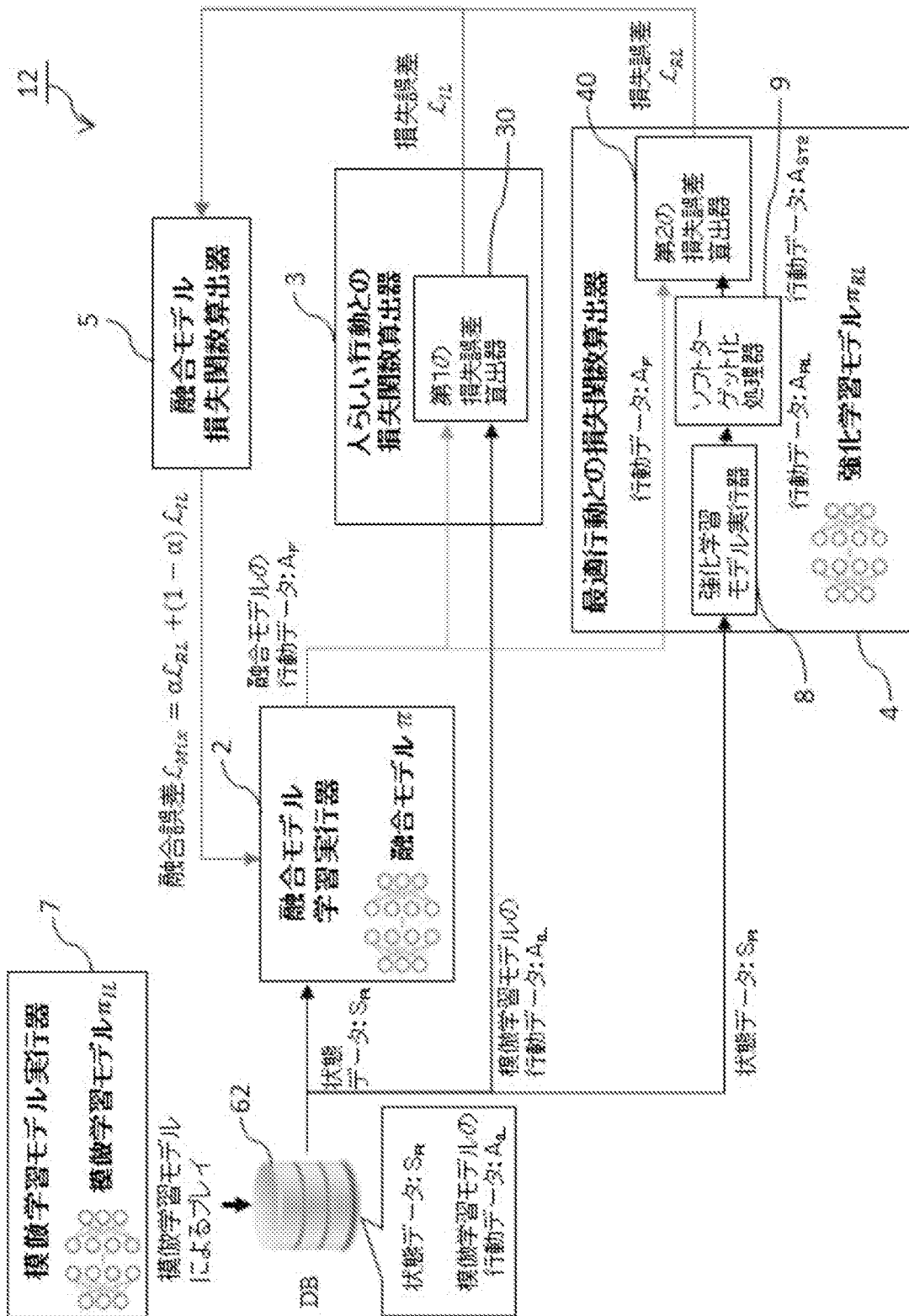
[図7]



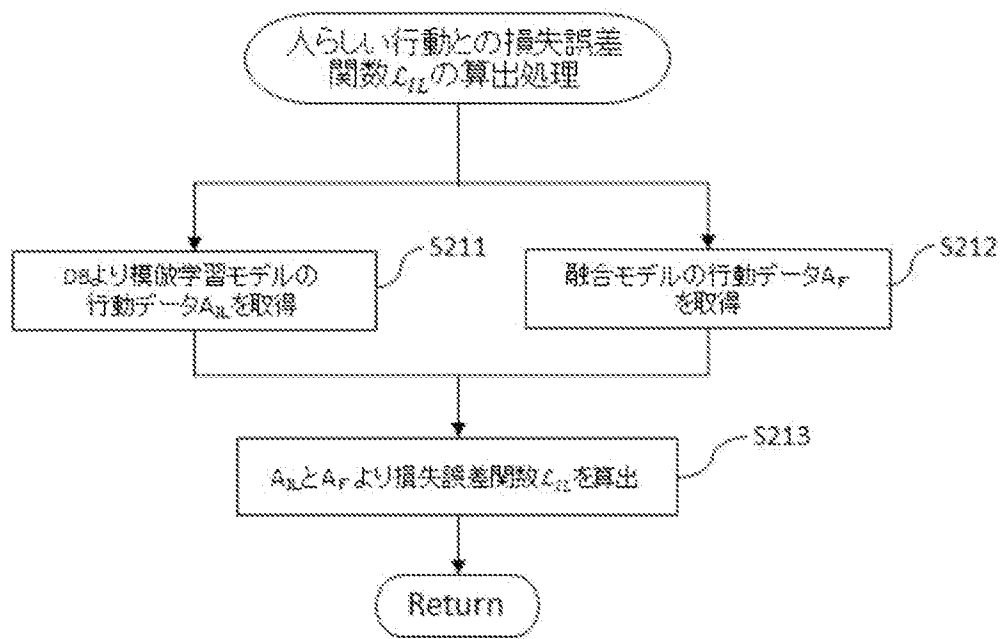
[図8]



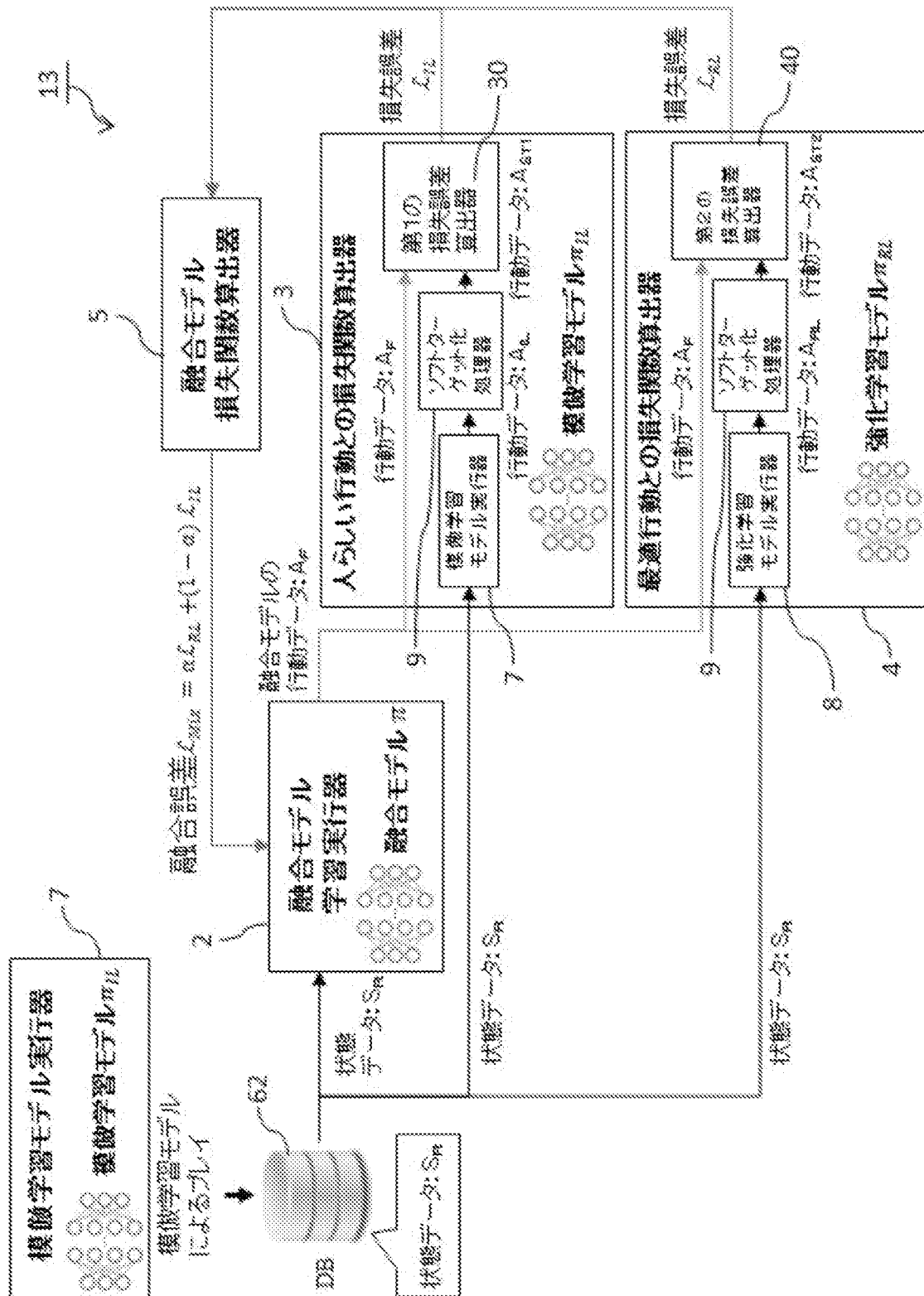
[図9]



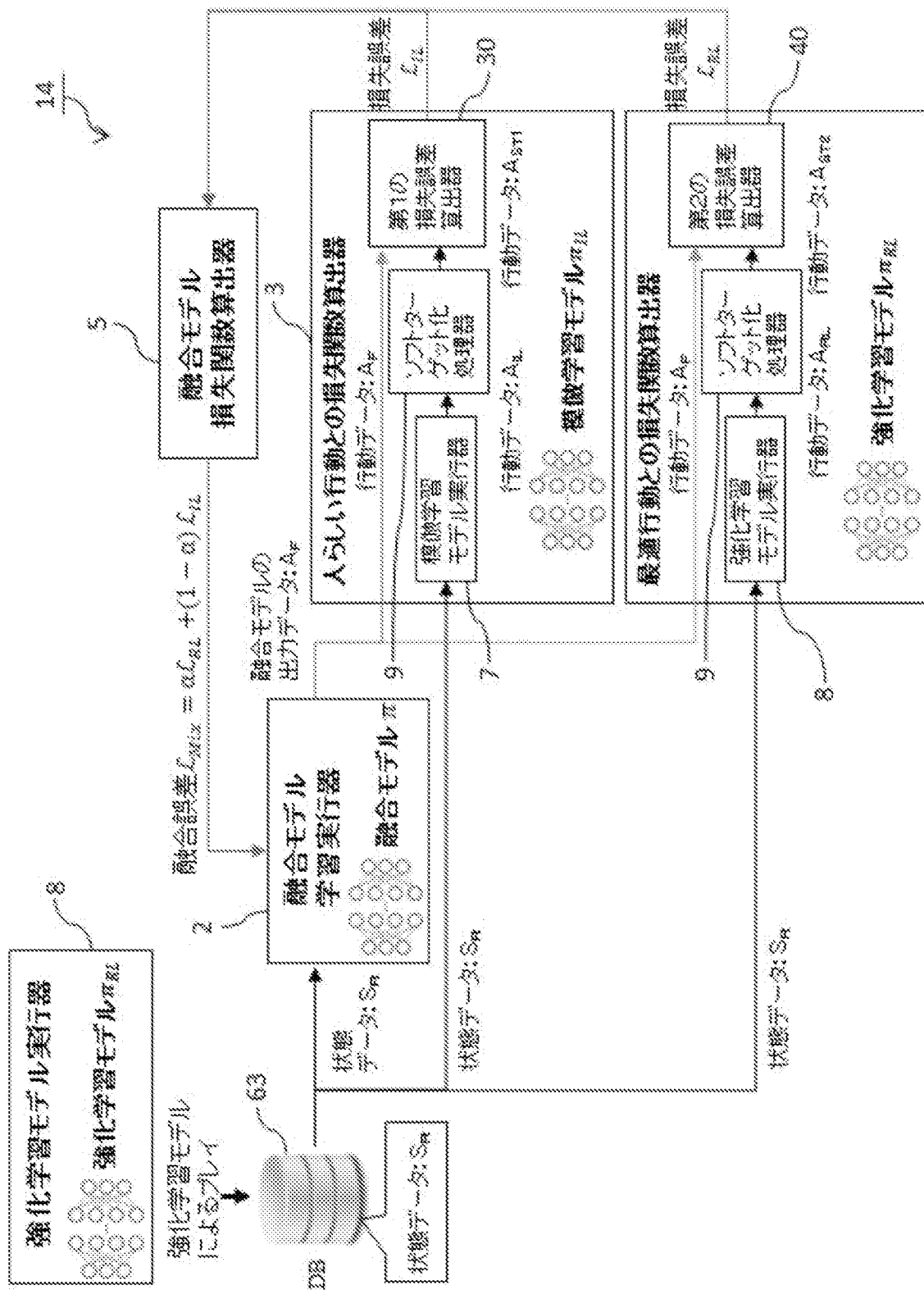
[図10]



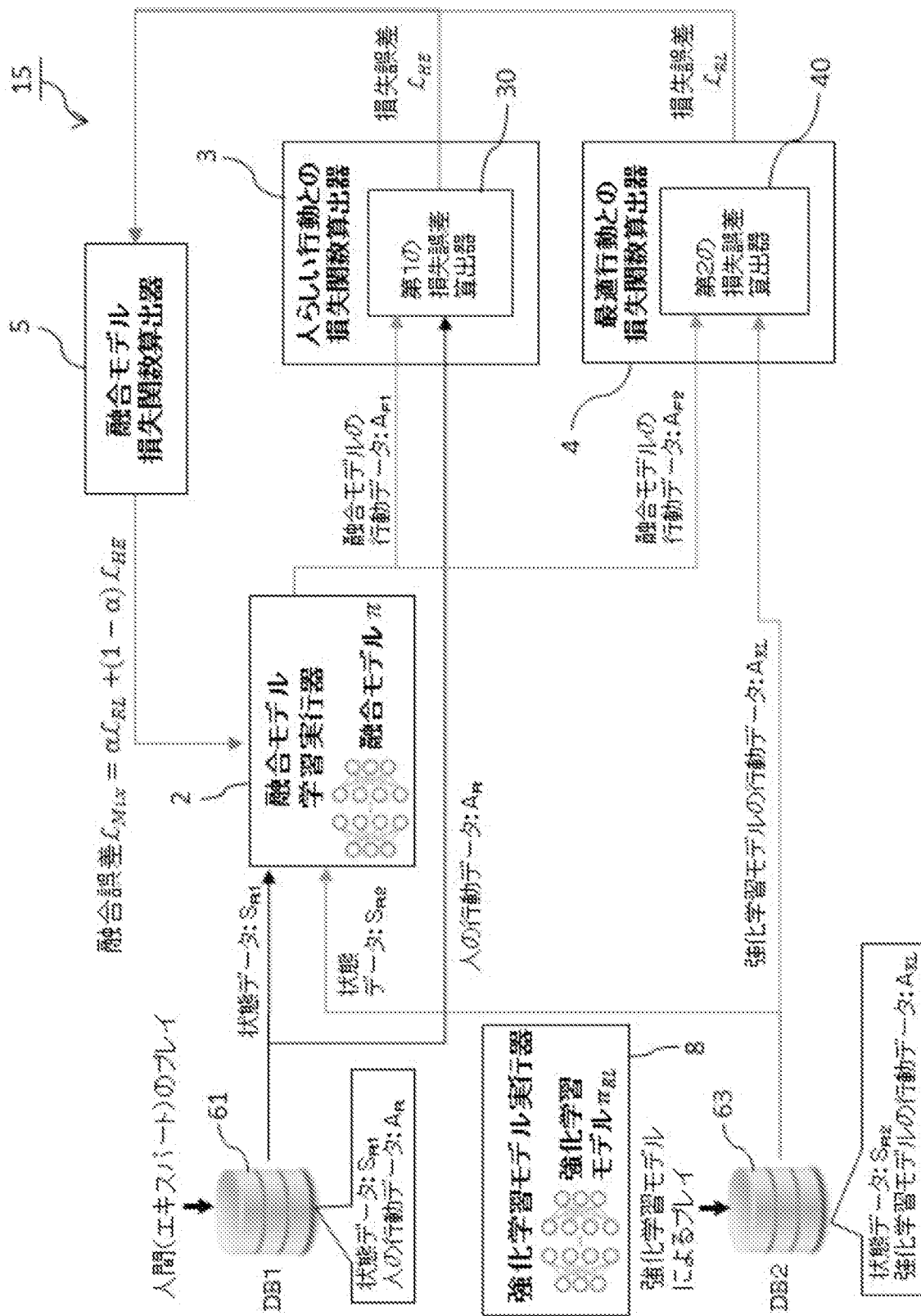
[図11]



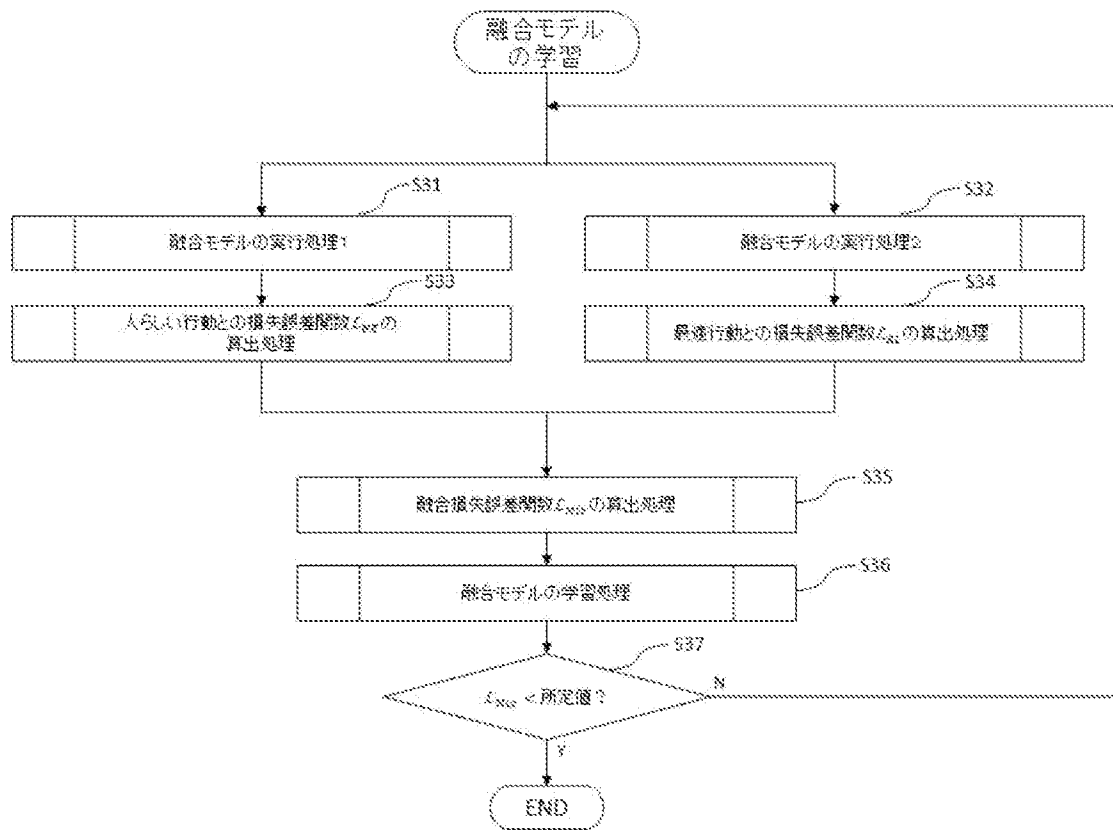
[図12]



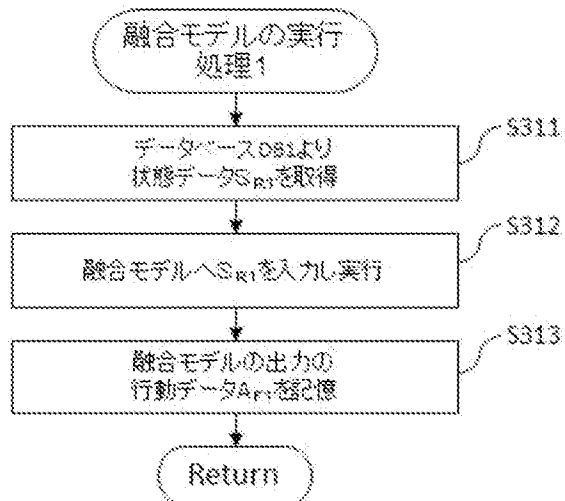
[図13]



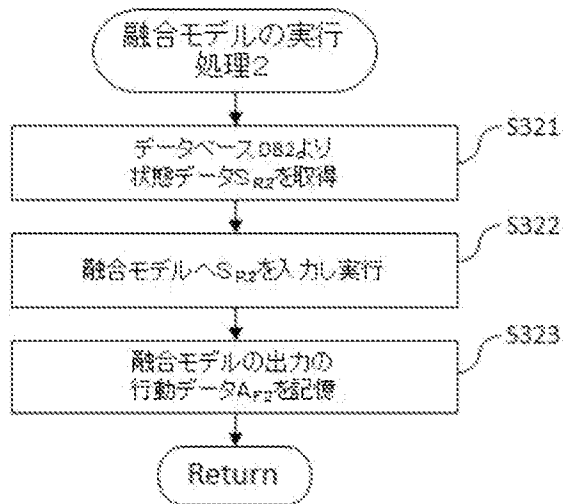
[図14]



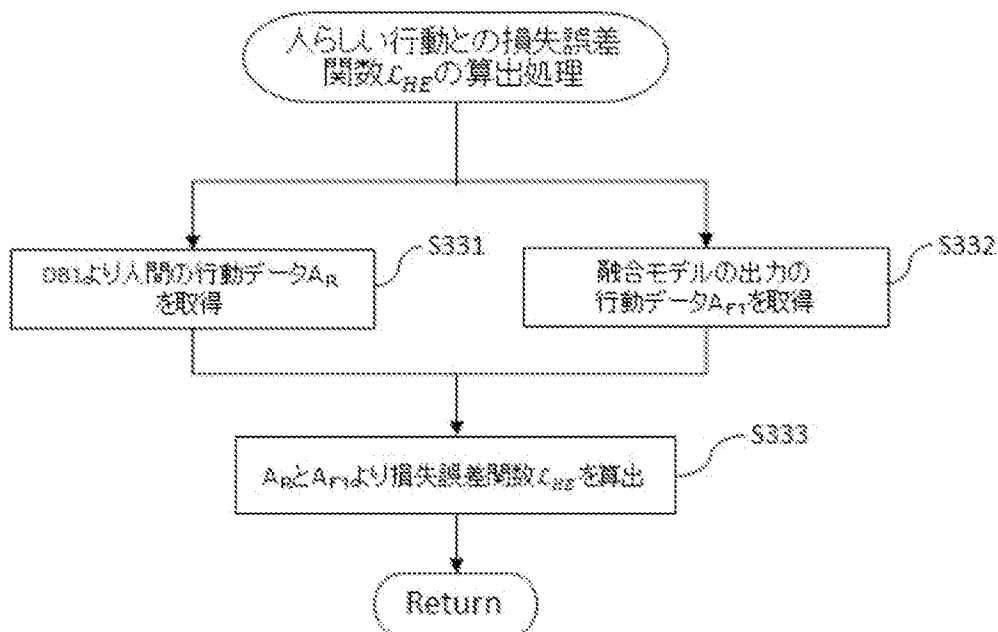
[図15]



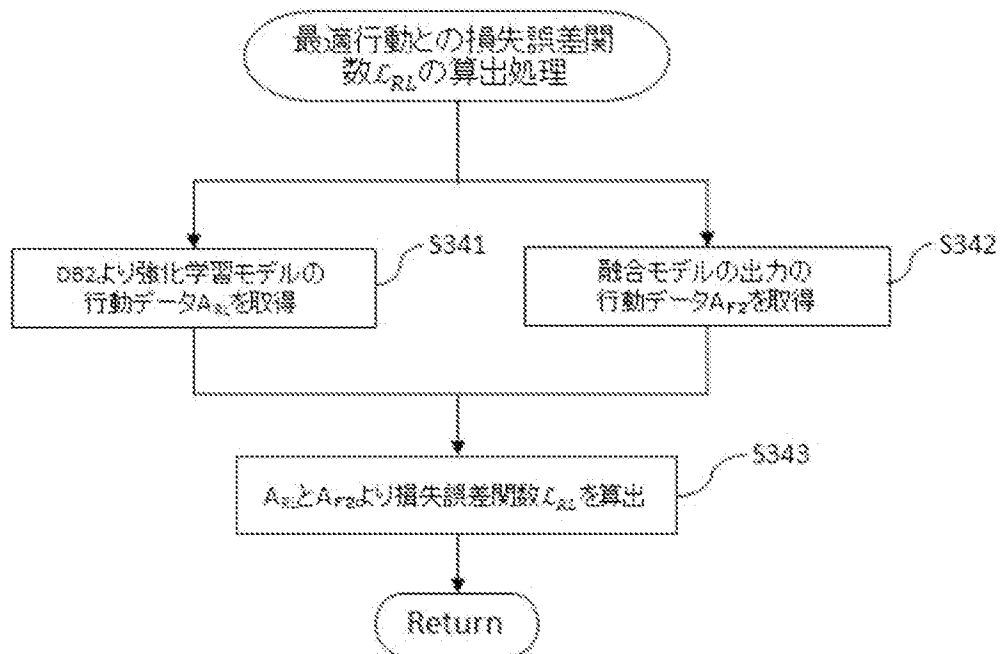
[図16]



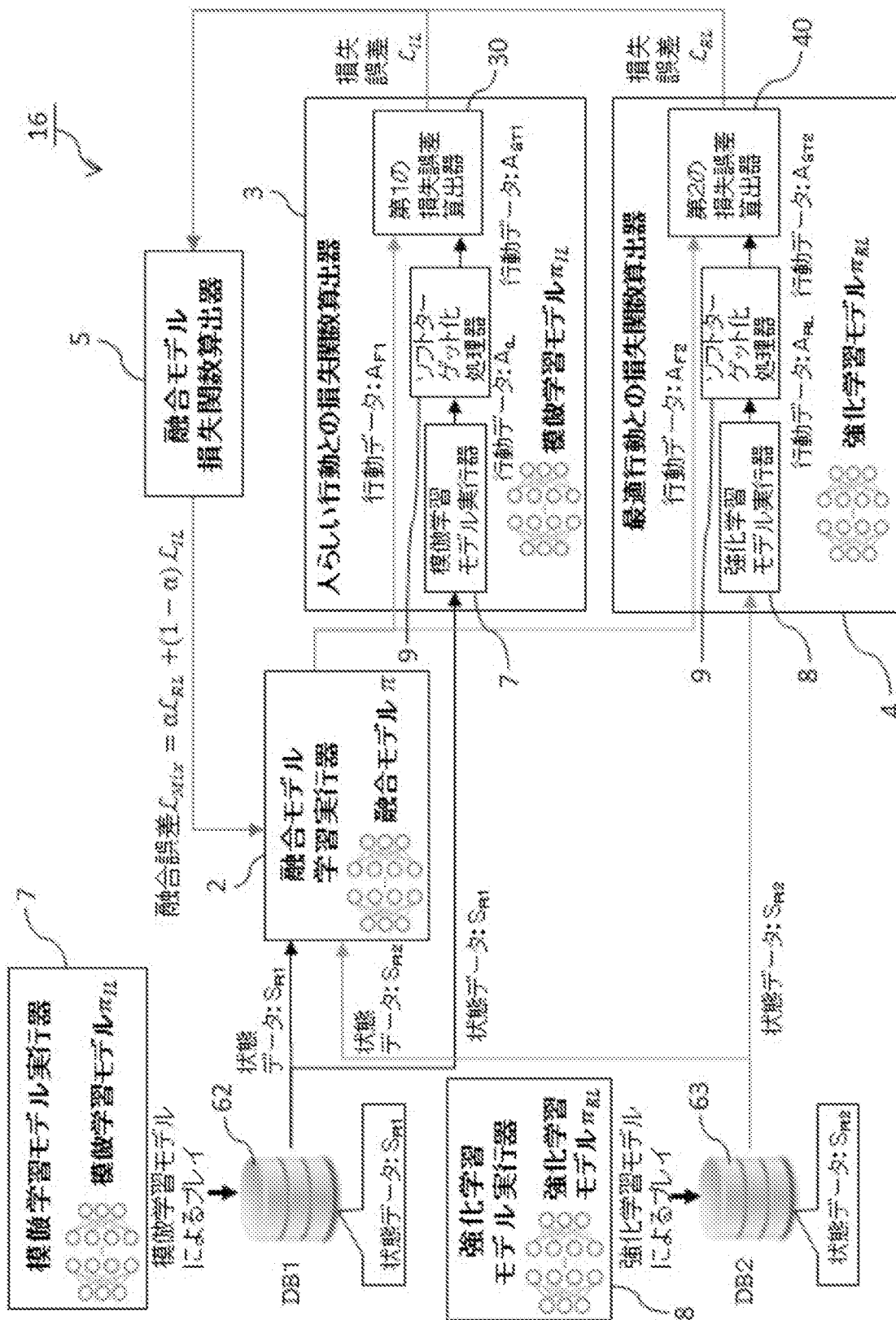
[図17]



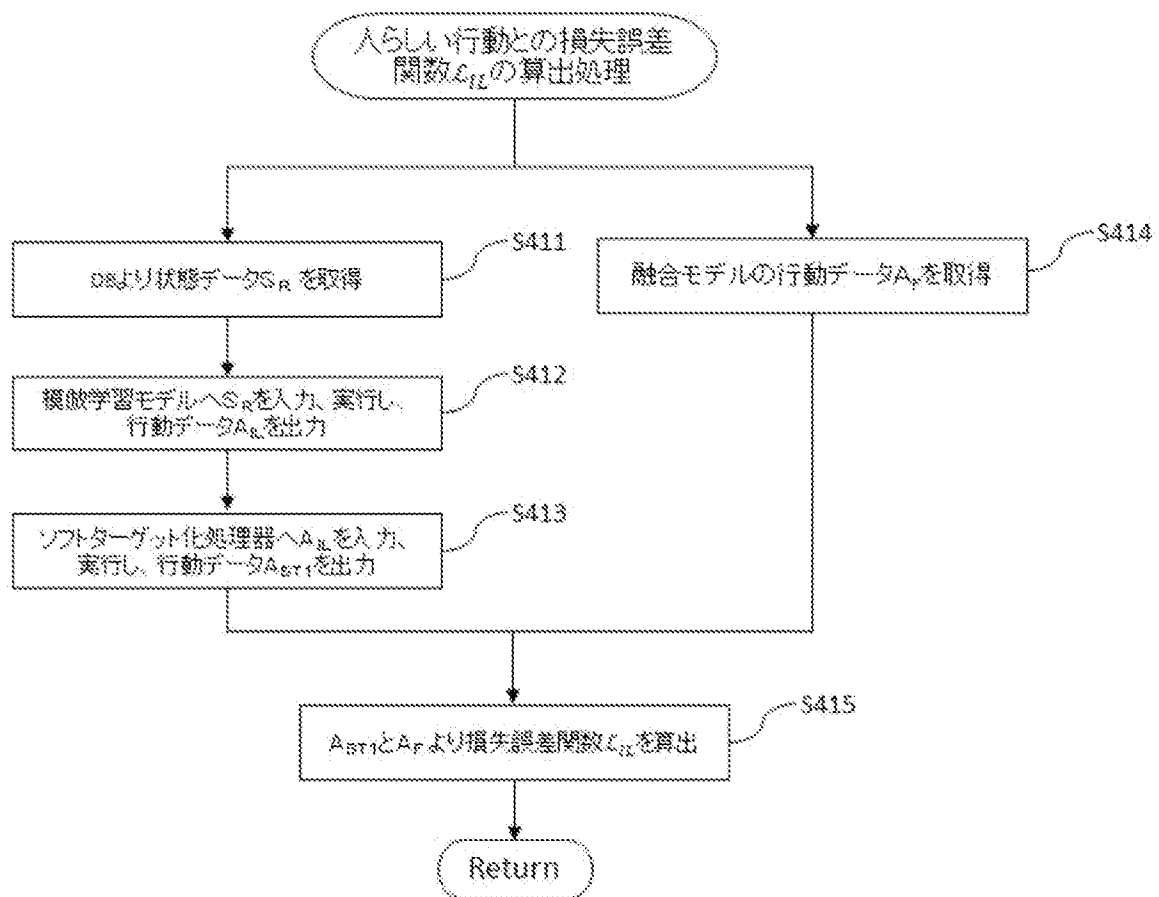
[図18]



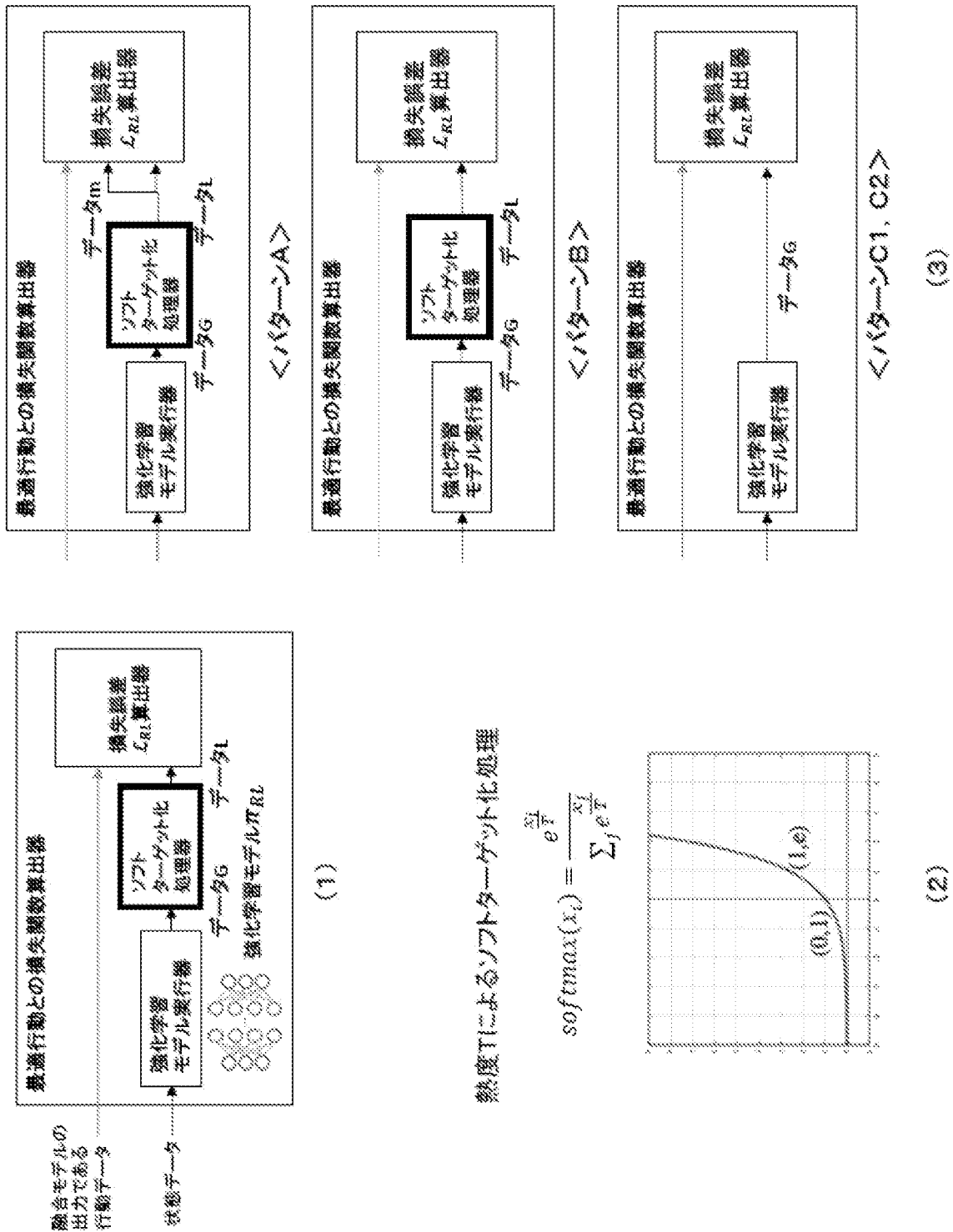
[図19]



[図20]

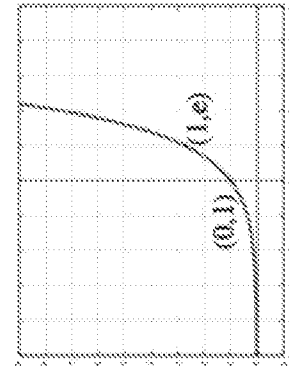


[図21]



熱度Tによるソフトターゲット化処理

$$\text{softmax}(x_i) = \frac{e^{\frac{x_i}{T}}}{\sum_j e^{\frac{x_j}{T}}}$$



(2)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2020/020624

A. CLASSIFICATION OF SUBJECT MATTER

Int.Cl. G06N20/00 (2019.01) i

FI: G06N20/00

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 1971-2020

Registered utility model specifications of Japan 1996-2020

Published registered utility model applications of Japan 1994-2020

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.
A	連欣瑜ほか, 強化学習と模倣学習の融合による人間らしいエージェントの構築, 電子情報通信学会技術研究報告, 2018.11.15, vol. 118, no. 316, pp. 45-50, ISSN 0913-5685, (LIAN, X. Y. et al., A human-like agent based on a hybrid of reinforcement and imitation learning, IEICE Technical Report), particularly, p. 48, "3. Proposed method"	1-12
A	MIYASHITA, Shohei et al., Developing game AI agent behaving like human by mixing reinforcement learning and supervised learning, Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2017), 2017.06.26, pp. 489-494, ISBN 978-1-5090-5504-3, particularly, pp. 490, 491, "III. Proposed method"	1-12

☐

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
25.06.2020Date of mailing of the international search report
07.07.2020Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC）） G06N 20/00(2019.01)i FI: G06N20/00														
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）） G06N 3/00-99/00														
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 <table border="0"> <tr> <td>日本国実用新案公報</td> <td>1922-1996年</td> </tr> <tr> <td>日本国公開実用新案公報</td> <td>1971-2020年</td> </tr> <tr> <td>日本国実用新案登録公報</td> <td>1996-2020年</td> </tr> <tr> <td>日本国登録実用新案公報</td> <td>1994-2020年</td> </tr> </table>			日本国実用新案公報	1922-1996年	日本国公開実用新案公報	1971-2020年	日本国実用新案登録公報	1996-2020年	日本国登録実用新案公報	1994-2020年				
日本国実用新案公報	1922-1996年													
日本国公開実用新案公報	1971-2020年													
日本国実用新案登録公報	1996-2020年													
日本国登録実用新案公報	1994-2020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												
A	連欣瑜ほか、強化学習と模倣学習の融合による人間らしいエージェントの構築、電子情報通信学会技術研究報告, 2018.11.15, 第118巻, 第316号, pp.45-50, ISSN 0913-5685 (LIAN, Xinyu et al., A Human-Like Agent Based on a Hybrid of Reinforcement and Imitation Learning, IEICE Technical Report) 特に48ページ「3. Proposed Method」	1-12												
A	MIYASHITA, Shohei et al., Developing Game AI Agent Behaving Like Human by Mixing Reinforcement Learning and Supervised Learning, Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2017), 2017.06.26, pp.489-494, ISBN 978-1-5090-5504-3 特に490-491ページ「III. Proposed Method」	1-12												
☐ 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。														
<table border="0"> <tr> <td>* 引用文献のカテゴリー</td> <td>“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抵触するものではなく、発明の原理又は理論の理解のために引用するもの</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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</td> <td>“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</td> </tr> <tr> <td>“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）</td> <td>“&” 同一パテントファミリー文献</td> </tr> <tr> <td>“O” 口頭による開示、使用、展示等に言及する文献</td> <td></td> </tr> <tr> <td>“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</td> <td></td> </tr> </table>			* 引用文献のカテゴリー	“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抵触するものではなく、発明の原理又は理論の理解のために引用するもの	“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	“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	“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	“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	“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）	“&” 同一パテントファミリー文献	“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	
* 引用文献のカテゴリー	“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抵触するものではなく、発明の原理又は理論の理解のために引用するもの													
“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	“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													
“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	“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													
“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）	“&” 同一パテントファミリー文献													
“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														
国際調査を完了した日 25.06.2020	国際調査報告の発送日 07.07.2020													
名称及びあて先 日本国特許庁(ISA/JP) 〒100-8915 日本国 東京都千代田区霞が関三丁目4番3号	権限のある職員（特許庁審査官） 北元 健太 5B 3856 電話番号 03-3581-1101 内線 3545													