(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
2024年8月15日(15.08.2024)

WIPO | PCT

(10) 国際公開番号
WO 2024/166331 A1

(51) 国際特許分類:
G06N 3/091 (2023.01) *G06N 20/00* (2019.01)

(21) 国際出願番号： PCT/JP2023/004458

(22) 国際出願日： 2023年2月9日(09.02.2023)

(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: 富士通株式会社 (**FUJITSU LIMITED**) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 Kanagawa (JP).

(72) 発明者: 園田 亮介 (**SONODA, Ryosuke**); 〒2118588 神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内 Kanagawa (JP).

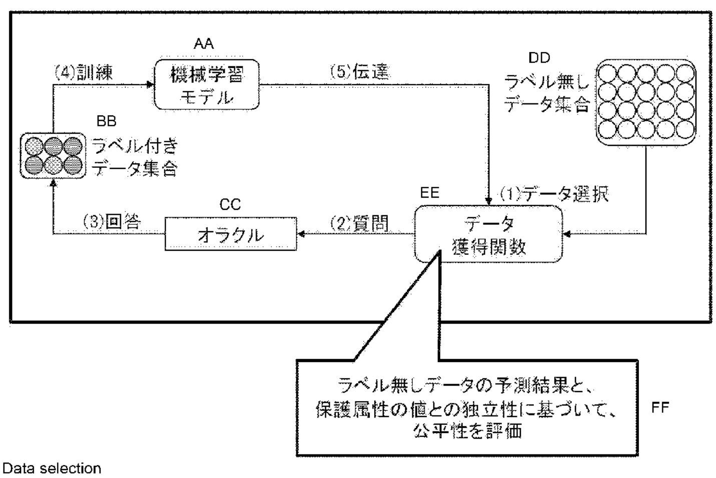
(74) 代理人: 弁理士法人太陽国際特許事務所(**TAIYO, NAKAJIMA & KATO**); 〒1600022 東京都新宿区新宿４丁目３番１７号 Tokyo (JP).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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.

(54) **Title:** MACHINE LEARNING PROGRAM, METHOD, AND DEVICE

(54) 発明の名称： 機械学習プログラム、方法、及び装置

［図3］



(1) Data selection
(2) Question
(3) Answer
(4) Training
(5) Transfer
AA Machine learning model
BB Labeled dataset
CC Oracle
DD Unlabeled dataset
EE Data acquisition function
FF Evaluate fairness based on independence between prediction result for unlabeled data and value of protected attribute

(57) **Abstract:** Provided is a machine learning device that: calculates the independence between a prediction result, which is obtained when inputting each of a plurality of pieces of unlabeled data into a machine learning model, and the value of a first attribute of each of the plurality of pieces of data; selects a first piece of data from the plurality of pieces of data on the basis of the independence; obtains a label for the first piece of data; and executes training of the machine learning model on the basis of the first piece of data and the label.

WO 2024/166331 A1

［続葉有］

(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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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：機械学習装置は、ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、複数のデータのそれぞれの第１の属性の値との独立性を算出し、独立性に基づいて、複数のデータから第１のデータを選択し、第１のデータのラベルを取得し、第１のデータとラベルとに基づいて機械学習モデルの訓練を実行する。

明　細　書

発明の名称 ： 機械学習プログラム、方法、及び装置

技術分野

[0001]　開示の技術は、機械学習プログラム、機械学習方法、及び機械学習装置に関する。

背景技術

[0002]　従来、公平性に配慮した機械学習モデルに関する技術が提案されている。例えば、分類器を学習するための訓練データと、訓練データに含まれる変数間の因果関係を表す因果グラフとを入力する学習装置が提案されている。この学習装置は、入力された訓練データと因果グラフとを用いて、所定の変数間の因果効果の平均が所定の範囲内にあり、かつ、因果効果の分散が所定の値以下である制約付き最適化問題を解くことにより分類器を学習する。

[0003]　また、例えば、少数属性のデータを人工的に増やすことによって、各入力データに対して公平な判断を行うための学習用データを生成する情報処理装置が提案されている。この情報処理装置は、機械学習モデルの学習に使用する第１の学習用データを保持し、学習用データの偏りに関する情報を取得し、偏りに関する情報に基づいて、学習用データに含まれるデータを用いて第２の学習用データを生成する。そして、この情報処理装置は、第１の学習用データ及び第２の学習用データを用いて機械学習モデルを学習する。

[0004]　また、例えば、ラベル付けされていないデータに、ラベルバイアスの量に従ってラベル付けするシステムが提案されている。このシステムは、選択バイアスの量及び特徴の希少性間の不一致に従って入力データをサンプリングし、サンプリングされてラベル付けされたデータと、追加のラベル付けされていないデータとを使用して分類器を訓練する。

[0005]　また、公平性に配慮した能動学習により機械学習モデルの訓練を実行する手法として、能動学習と半教師あり学習とを統合した手法が提案されている。この手法は、ラベル付けされていない最も価値のあるデータを選択し、ラ

ベル付けのためにエキスパートシステムに送信する。そして、この手法は、ラベル付けされていないデータとラベル付けされているデータとの間の接続を確立し、ラベル付けされていないデータの固有の情報を使用してモデルを改善し、疑似ラベルをそれらのサンプルに割り当てる。

先行技術文献

特許文献

[0006] 特許文献1：国際公開第２０２１／０８４６０９号公報

特許文献2：国際公開第２０２２／１２３９０７号公報

特許文献3：米国特許出願公開第２０２０／０３７２４０６号明細書

非特許文献

[0007] 非特許文献1：Quan Ren, Hongbing Zhang, Dailu Zhang, Xiang Zhao, Lizhi Yan, Jianwen Rui, Fanxin Zeng, Xinyi Zhu, "A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes," Expert Systems with Applications, Volume 202, 15 September 2022, 117278.

発明の概要

発明が解決しようとする課題

[0008] しかしながら、従来の公平性に配慮した能動学習により機械学習モデルの訓練を実行する学習方法（以下、「公平能動学習」という）では、ラベル付けの対象となるデータを選択する処理においても機械学習モデルの訓練が必要であり、処理負荷が高いという問題がある。

[0009] 一つの側面として、開示の技術は、公平能動学習の処理負荷を低減することを目的とする。

課題を解決するための手段

[0010] 一つの態様として、開示の技術は、ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、前記複数のデータのそれぞれの第１の属性の値との独立性を算出する。そして、開示の技術

は、前記独立性に基づいて、前記複数のデータから第１のデータを選択し、前記第１のデータのラベルを取得し、前記第１のデータと前記ラベルとに基づいて前記機械学習モデルの訓練を実行する。

発明の効果

[0011]　一つの側面として、開示の技術は、公平能動学習の処理負荷を低減することができる、という効果を有する。

図面の簡単な説明

[0012] [図1]能動学習を説明するための図である。

[図2]従来の公平能動学習を説明するための図である。

[図3]本実施形態における公平能動学習を説明するための図である。

[図4]本実施形態に係る機械学習装置の機能ブロック図である。

[図5]精度改善度合いの一例を説明するための図である。

[図6]機械学習装置として機能するコンピュータの概略構成を示すブロック図である。

[図7]機械学習処理の一例を示すフローチャートである。

[図8]機械学習モデルの予測精度、公平性、及び公平能動学習の実行時間について、本手法と比較手法との比較を示す図である。

発明を実施するための形態

[0013]　以下、図面を参照して、開示の技術に係る実施形態の一例を説明する。

[0014]　実施形態の詳細を説明する前に、機械学習モデルの公平能動学習及びその課題について説明する。

[0015]　まず、機械学習モデルの訓練とは、データのラベル（結果変数）と特徴量（説明変数）との関係を学習し、その関係を近似するパラメータを特定することである。また、機械学習における公平性とは、機械学習モデルによる予測結果に、意思決定における個人や集団の先天的又は後天的な特性（以下、「保護属性」という）に基づく偏りや差別がないことである。機械学習モデルの社会実装では、ローン審査における性差、顔認識における人種差、病気診断における年齢差等、集団に基づく公平性及び格差の指標の改善が求めら

れることが多い。そのため、公平な予測結果が得られるように機械学習モデルを訓練することが重要である。また、機械学習モデルの公平性と予測精度とはトレードオフの関係にあるため、これらのバランスをとることも重要である。

[0016] 教師あり学習により機械学習モデルを訓練する場合、正解を示すラベルが付与されたデータ（以下、「ラベル付きデータ」という）が必要である。ラベル付きデータは、１つ以上のラベルでタグ付けされたデータである。ラベルは通常、人間、その他情報源であるオラクルによってタグ付けされる。十分な数のラベル付きデータがない場合、機械学習モデルの公平性及び予測精度を十分に改善することができない。しかし、ラベル付きデータは、ラベル付けがされていないデータ（以下、「ラベル無しデータ」という）よりも収集コストが高い。

[0017] また、能動学習とは、質問によって機械学習モデルを改善する対話的な機械学習手法である。具体的には、能動学習を実行する情報処理装置は、図１に示すように、（１）データ選択、（２）質問、（３）回答、（４）訓練、及び（５）伝達の処理を実行する。

[0018] より具体的には、情報処理装置は、「（１）データ選択」の処理として、ラベル無しデータ集合に含まれる各データについて、データ獲得関数を算出し、データ獲得関数に基づいて、機械学習モデルの訓練に有用なデータを優先して選択する。データ獲得関数は、現在の機械学習モデルのパラメータ等の情報を用いて、各データに対する機械学習モデルによる予測の曖昧さ、ラベル無しデータ集合に対する各データの代表性等を表す指標である。また、情報処理装置は、「（２）質問」の処理として、選択したデータのラベルをオラクルへ問い合わせる。また、情報処理装置は、「（３）回答」の処理として、オラクルからの回答であるラベルを取得し、取得したラベルを選択したデータに付与してラベル付きデータとし、ラベル付きデータ集合に追加する。また、情報処理装置は、「（４）訓練」の処理として、ラベル付きデータ集合を用いて機械学習モデルを訓練する。また、情報処理装置は、「（５

）伝達」の処理として、訓練後の機械学習モデルのパラメータ等の情報を（１）データ選択の処理へ伝達する。

[0019] 能動学習では、上記（１）～（５）の処理が繰り返し実行されることにより、ラベル無しデータのうち、機械学習モデルの訓練に有用なデータから優先的にラベル付けが行われるため、効果的にラベル付きデータを収集することができる。

[0020] なお、図１において、丸は各データを表し、白丸はラベル無しデータ、ハッチング付きの丸はラベル付きデータを表し、ハッチングの相違はラベルの相違を表す。以下の図２及び図３においても同様である。

[0021] 通常の能動学習では、（１）データ選択において、機械学習モデルの予測精度の改善に役立つデータが選択され、公平性が考慮されない。そのため、ラベル付きデータが追加され、機械学習モデルの訓練が進むほど、公平性の悪化を招く場合がある。例えば、顔表情認識のための機械学習モデルにおいて、ラベル無しデータ集合から特定の人種のデータばかりが選択されてしまう場合が考えられる。

[0022] そこで、従来の公平能動学習では、（１）データ選択の処理で、公平性と予測精度とのトレードオフを考慮してデータを選択する。具体的には、図２に示すように、従来の公平能動学習では、ラベル無しデータ集合の一部をラベル無し検証データ集合とし、残りをラベル無し候補データ集合とする。従来の公平能動学習を実行する情報処理装置は、図２に示す（Ａ）仮質問、（Ｂ）仮回答、（Ｃ）訓練、（Ｄ）評価、及び（Ｅ）選択の処理を実行することで、各候補データについて、検証データにおける機械学習モデルの不公平性改善度合いを推定する。

[0023] より具体的には、情報処理装置は、「（Ａ）仮質問」の処理として、データに対する仮ラベルを出力するラベル付けモデルに各候補データを入力する。また、情報処理装置は、「（Ｂ）仮回答」の処理として、ラベル付けモデルから出力された仮ラベルを取得し、取得した仮ラベルを各候補データに付与して、仮ラベル付き候補データ集合とする。また、情報処理装置は、「（

C）訓練」の処理として、仮ラベル付き候補データ集合を用いて機械学習モデルを訓練する。また、情報処理装置は、「（D）評価」の処理として、ラベル無し検証データ集合を用いて、訓練の前後での機械学習モデルの不公平性の差分を考慮して、各仮ラベル付き候補データの公平性を評価する。また、情報処理装置は、ラベル無し検証データ集合を用いて機械学習モデルの予測精度を評価する。また、情報処理装置は、「（E）選択」の処理として、公平性と予測精度とのトレードオフを考慮した指標に基づいて、最も値の良い候補データを選択する。

[0024] 従来の公平能動学習では、「（1）データ選択」において、機械学習モデルの訓練が実行されるため、機械学習モデルの決定境界を考慮して、予測精度が改善するようなデータが選択される。また、検証データを用いた評価により、公平性も改善するようなデータが選択される。しかし、機械学習モデルの訓練は処理負荷が高く、効率的にデータ選択を実行することができないという課題がある。特に、機械学習モデルが深層学習モデル、非線形モデル等の複雑なモデルの場合には、現実的な実行時間で公平能動学習を適用することが困難である。

[0025] そこで、本実施形態では、図３に示すように、ラベル無しデータに対するモデルの予測結果と、保護属性の値との独立性に基づいて、機械学習モデルの訓練を要することなく、各候補データの公平性を評価する。

[0026] また、上記の課題を解決するために容易に思いつく手段として、各候補データに対する機械学習モデルの予測結果に基づいて、各候補データの不公平性度合いを推定し、ラベル無し候補データ集合内から不公平性度合いが低いデータを選択することが考えられる。しかし、この場合、選択されるデータは、検証データへの影響が考慮されていないため、外れ値又は同じようなデータが選択され易い等、データの代表性が低くなる。そこで、さらに、本実施形態では、ラベル無しデータに対するモデルの予測結果として、候補データが与えられた場合の検証データの予測結果を用いる。以下、本実施形態に係る機械学習装置について説明する。

[0027]　図４に示すように、機械学習装置１０には、ラベル付きデータ集合２０及びラベル無しデータ集合２２が入力される。ラベル付きデータ集合２０に含まれるラベル付きデータの数は、ラベル無しデータ集合２２に含まれるラベル無しデータの数に比べて非常に少ないものとする。機械学習装置１０は、ラベル無しデータ集合２２から、公平性に配慮したデータを選択して機械学習モデル２４の訓練を実行する。すなわち、機械学習装置１０は、公平能動学習を実行する。

[0028]　機械学習装置１０は、機能的には、図４に示すように、制御部１１を含む。制御部１１は、さらに、訓練部１２と、算出部１４と、選択部１６と、取得部１８とを含む。また、機械学習装置１０の所定の記憶領域には、機械学習モデル２４が記憶される。

[0029]　訓練部１２は、ラベル付きデータ集合２０に含まれる複数のラベル付きデータを訓練データとして用いて、機械学習モデル２４の訓練を実行する。また、後述するように、本実施形態では、取得部１８により、ラベル付きデータ集合２０に新たなラベル付きデータが追加される。ラベル付きデータ集合２０に新たなラベル付きデータが追加された場合には、訓練部１２は、当初のラベル付きデータと、追加されたラベル付きデータとを用いて、機械学習モデル２４の訓練を実行する。

[0030]　算出部１４は、ラベル無しデータ集合２２に含まれる複数のラベル無しデータのそれぞれを機械学習モデル２４に入力した場合の予測結果（以下、「予測ラベル」という）と、複数のデータのそれぞれの保護属性の値との独立性を算出する。保護属性は、開示の技術の「第１の属性」の一例である。算出部１４は、独立性として、予測ラベルと保護属性の値との相互情報量を算出する。相互情報量は、２つの変数が依存しているか否かを定量的に表す指標であり、２つの変数がお互いに完全に独立している場合は、相互情報量は０となる。すなわち、予測ラベルと保護属性の値との相互情報量が０の場合、機械学習モデル２４は完全に公平であると言える。したがって、この相互情報量が最も小さいラベル無しデータが最も公平なデータであると言える。

[0031] 本実施形態では、算出部１４は、各候補データで条件付けられた、検証データについての予測ラベルと保護属性の値との相互情報量に基づいて、機械学習モデル２４の不公平性改善度合いを算出する。具体的には、算出部１４は、ラベル無しデータ集合２２に含まれる複数のラベル無しデータの一部を検証データ、検証データ以外を候補データとして設定する。算出部１４は、各検証データの予測ラベルと保護属性の値との独立性であって、各候補データの予測ラベルと保護属性の値との独立性に条件付けられた独立性を算出する。

[0032] より具体的には、算出部１４は、検証データｖの予測ラベルY_vと、検証データｖにおける保護属性の値S_vとの相互情報量Ｉの、候補データｕが与えられる前後での差分を、候補データｕの不公平性改善度合いF_uとして、下記（１）式により算出する。

[0033] [数1]

$$F_u = \sum_v \{I(Y_v; S_v) - I(Y_v; S_v | Y_u, S_u)\} \qquad (1)$$

[0034] （１）式において、Y_uは、候補データｕの予測ラベル、S_uは、候補データｕにおける保護属性の値である。（１）式右辺のΣ内の第１項は、候補データｕが与えられる前の検証データｖの相互情報量であり、第２項は、候補データｕが与えられた後の検証データｖの相互情報量である。（１）式の場合、不公平性改善度合いF_uの値が大きい候補データｕほど、不公平性の改善度合いが高いことを表す。

[0035] なお、算出部１４は、（１）式右辺のΣ内の第２項について、下記のように算出する。まず、算出部１４は、第２項を下記（２）式のように変換する。（２）式において、Ｈ（Ｘ）は、Ｘのエントロピーである。

[0036] [数2]

$$\begin{aligned} & I(Y_v; S_v | Y_u, S_u) \\ & = H(Y_v | Y_u, S_u) - H(Y_v | S_v, Y_u, S_u) \\ & = \{H(Y_v, Y_u, S_u) - H(Y_u, S_u)\} \\ & - \{H(Y_v, S_v, Y_u, S_u) - H(S_v, Y_u, S_u)\} \end{aligned} \qquad (2)$$

[0037] 次に、算出部１４は、モンテカルロドロップアウトにより、各エントロピーに対応する確率分布を近似する。算出部１４は、機械学習モデル２４のパラメータと、機械学習モデル２４の予測ラベルの分布とが条件付き独立であると仮定し、下記（３）式により、Y_iの確率Ｐ（Y_i）を算出する。ここでは、Y_i＝｛Y_v，S_v，Y_u，S_u｝である。

[0038] [数3]

$$P(Y_i) = P(Y_i|\boldsymbol{\theta}) \approx \frac{1}{M}\sum_{m=1}^{M} P(Y_i|\boldsymbol{\theta}) \qquad (3)$$

[0039] （３）式において、θは、機械学習モデル２４のパラメータであり、Ｍは、モンテカルロサンプリング回数である。算出部１４は、（３）式の確率分布を用いて、（２）式の相互情報量を算出する。

[0040] また、算出部１４は、各候補データを機械学習モデル２４に入力した場合の予測結果の不確実性に基づく、各候補データによる機械学習モデル２４の予測精度の改善度合いを、各候補データについて算出する。例えば、機械学習モデル２４の決定境界付近に位置するデータは、機械学習モデル２４にとって判断が難しいデータと言えるため、このようなデータを訓練データとして用いることで、機械学習モデル２４の予測精度の改善が図られる。例えば、機械学習モデル２４において、図５に示すように、特徴空間で決定境界が規定されているとする。なお、図５の例では、線形モデルの２値分類の例を示しており、各丸は各データの特徴量を表す。この場合、決定境界付近のデータ（例えば、図５中の網点の丸で示すデータ）は、どちらのラベルに属するか不安定であり、機械学習モデル２４の精度改善に有用である。そこで、算出部１４は、候補データが機械学習モデルの決定境界に近いほど高くなる精度改善度合いを算出する。例えば、算出部１４は、下記（４）式に示すように、候補データｕの予測ラベルY_uの不確かさを示すエントロピーを、精度改善度合いA_uとして算出する。（４）式において、Ｙは、機械学習モデル２４の予測ラベルの集合である。

[0041] [数4]

$$A_u = H(Y_u|u) = -\sum_{Y_u \in Y} P(Y_u|u) \log P(Y_u|u) \quad (4)$$

[0042] また、算出部１４は、例えば、下記（５）式に示すように、不公平性改善度合いＦ$_u$と、精度改善度合いＡ$_u$と、不公平性改善度合いＦ$_u$と精度改善度合いＡ$_u$とのトレードオフを表す係数αとで表される、各候補データｕについての評価値Ｅ$_u$を算出する。

Ｅ$_u$＝α×Ｆ$_u$＋（１－α）×Ａ$_u$ （５）

αは、０～１の値（例えば、０．６）であり、不公平性改善度合いＦ$_u$と精度改善度合いＡ$_u$とのどちらをどの程度優先させるかを規定する係数である。

[0043] 選択部１６は、算出部１４により算出された各候補データｕについての評価値Ｅ$_u$に基づいて、複数の候補データｕから、ラベル付けの対象となる対象データを選択する。対象データは、開示の技術の「第１のデータ」の一例である。例えば、選択部１６は、評価値Ｅ$_u$が最も高い候補データｕを選択してもよいし、評価値Ｅ$_u$が所定値以上の候補データｕを選択してもよいし、評価値Ｅ$_u$が上位所定個の候補データｕを選択してもよい。

[0044] 取得部１８は、選択部１６により選択された対象データのラベルを、人間、その他情報源であるオラクルへ問い合わせ、オラクルからの回答であるラベルを取得する。取得部１８は、取得したラベルを対象データに付与してラベル付きデータとし、ラベル付きデータ集合２０に追加する。これにより、上述したように、訓練部１２が、追加されたラベル付きデータも用いて、機械学習モデル２４の訓練を実行する。

[0045] 機械学習装置１０は、例えば図６に示すコンピュータ４０で実現されてよい。コンピュータ４０は、ＣＰＵ（Central Processing Unit）４１と、ＧＰＵ（Graphics Processing Unit）４２と、一時記憶領域としてのメモリ４３と、不揮発性の記憶装置４４とを備える。また、コンピュータ４０は、入力装置、表示装置等の入出力装置４５と、記憶媒体４９に対するデータの読み

込み及び書き込みを制御するＲ／Ｗ（Read/Write）装置４６とを備える。また、コンピュータ４０は、インターネット等のネットワークに接続される通信Ｉ／Ｆ（Interface）４７を備える。ＣＰＵ４１、ＧＰＵ４２、メモリ４３、記憶装置４４、入出力装置４５、Ｒ／Ｗ装置４６、及び通信Ｉ／Ｆ４７は、バス４８を介して互いに接続される。

[0046]　記憶装置４４は、例えば、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、フラッシュメモリ等である。記憶媒体としての記憶装置４４には、コンピュータ４０を、機械学習装置１０として機能させるための機械学習プログラム５０が記憶される。機械学習プログラム５０は、訓練プロセス制御命令５２と、算出プロセス制御命令５４と、選択プロセス制御命令５６と、取得プロセス制御命令５８とを有する。また、記憶装置４４は、機械学習モデル２４を構成する情報が記憶される情報記憶領域６０を有する。

[0047]　ＣＰＵ４１は、機械学習プログラム５０を記憶装置４４から読み出してメモリ４３に展開し、機械学習プログラム５０が有する制御命令を順次実行する。ＣＰＵ４１は、訓練プロセス制御命令５２を実行することで、図４に示す訓練部１２として動作する。また、ＣＰＵ４１は、算出プロセス制御命令５４を実行することで、図４に示す算出部１４として動作する。また、ＣＰＵ４１は、選択プロセス制御命令５６を実行することで、図４に示す選択部１６として動作する。また、ＣＰＵ４１は、取得プロセス制御命令５８を実行することで、図４に示す取得部１８として動作する。また、ＣＰＵ４１は、情報記憶領域６０から情報を読み出して、機械学習モデル２４をメモリ４３に展開する。これにより、機械学習プログラム５０を実行したコンピュータ４０が、機械学習装置１０として機能することになる。なお、プログラムを実行するＣＰＵ４１はハードウェアである。また、プログラムの一部は、ＧＰＵ６２により実行されてもよい。

[0048]　なお、機械学習プログラム５０により実現される機能は、例えば半導体集積回路、より詳しくはＡＳＩＣ（Application Specific Integrated Circuit）、ＦＰＧＡ（Field-Programmable Gate Array）等で実現されてもよい。

[0049]　次に、本実施形態に係る機械学習装置１０の動作について説明する。機械学習モデル２４に対する公平能動学習が指示されると、機械学習装置１０が、図７に示す機械学習処理を実行する。なお、機械学習処理は、開示の技術の機械学習方法の一例である。

[0050]　ステップＳ１０で、訓練部１２が、ラベル付きデータ集合２０を取得して、ラベル付きデータを訓練データとして用いて、機械学習モデル２４の訓練を実行する。次に、ステップＳ１２で、訓練部１２が、公平能動学習の終了条件を満たすか否かを判定する。終了条件は、例えば、ラベル付きデータ集合２０に新たに追加されたデータ数が所定数を超えた場合としてよい。終了条件を満たさない場合には、ステップＳ１４へ移行する。

[0051]　ステップＳ１４では、算出部１４が、ラベル無しデータ集合２２に含まれる複数のラベル無しデータの一部を検証データ、残りを候補データとして設定する。次に、ステップＳ１６で、算出部１４が、例えば（１）式に示す、検証データｖの予測ラベル$Ｙ_{ｖ}$と、検証データｖにおける保護属性の値$Ｓ_{ｖ}$との相互情報量Ｉの、候補データｕが与えられる前後での差分を、各候補データｕの不公平性改善度合い$Ｆ_{ｕ}$として算出する。

[0052]　次に、ステップＳ１８で、算出部１４が、例えば（４）式に示す、各候補データｕの予測ラベル$Ｙ_{ｕ}$の不確かさを示すエントロピーを、精度改善度合い$Ａ_{ｕ}$として算出する。次に、ステップＳ２０で、算出部１４が、例えば（５）式に示す、不公平性改善度合い$Ｆ_{ｕ}$と、精度改善度合い$Ａ_{ｕ}$と、不公平性改善度合い$Ｆ_{ｕ}$と精度改善度合い$Ａ_{ｕ}$とのトレードオフを表す係数αとで表される、各候補データｕについての評価値$Ｅ_{ｕ}$を算出する。

[0053]　次に、ステップＳ２２で、選択部１６が、各候補データｕについての評価値$Ｅ_{ｕ}$に基づいて、複数の候補データｕから、ラベル付けの対象となる対象データを選択する。次に、ステップＳ２４で、取得部１８が、対象データのラベルをオラクルへ問い合わせ、オラクルからの回答であるラベルを取得する。次に、ステップＳ２６で、取得部１８が、取得したラベルを対象データに付与してラベル付きデータとし、ラベル付きデータ集合２０に追加すると共

に、対象データとなった候補データをラベル無しデータ集合２２から削除し、ステップＳ１０に戻る。

[0054] ステップＳ１０に戻ると、訓練部１２が、追加されたラベル付きデータも用いて、機械学習モデル２４の訓練を実行する。次に、ステップＳ１２で、公平能動学習の終了条件を満たすと判定されると、ステップＳ２８へ移行する。ステップＳ２８では、訓練部１２が、公平能動学習による訓練済みの機械学習モデルを出力し、機械学習処理は終了する。

[0055] 以上説明したように、本実施形態に係る機械学習装置は、複数のラベル無しデータのそれぞれを機械学習モデルに入力した場合の予測結果と、複数のラベル無しデータのそれぞれの保護属性の値との独立性を算出する。また、機械学習装置は、算出した独立性に基づいて、複数のラベル無しデータから対象データを選択し、オラクルへ問い合わせて対象データのラベルを取得し、対象データと取得したラベルとに基づいて機械学習モデルの訓練を実行する。すなわち、本実施形態に係る機械学習装置は、ラベル無しデータの公平性を情報理論的アプローチに基づいて評価し、機械学習モデルの再学習を行うことなく、ラベル付けの対象となるデータを選択する。これにより、本実施形態に係る機械学習装置は、公平能動学習の処理負荷を低減することができる。

[0056] また、本実施形態に係る機械学習装置は、各候補データについて、予測結果と保護属性の値との独立性として相互情報量を用いた不公平性改善度合いを算出すると共に、候補データの不確かさに基づいて、機械学習モデルの精度改善度合いを算出する。そして、機械学習装置は、不公平性改善度合いと精度改善度合いとのトレードオフを考慮した評価値に基づいて対象データを選択する。これにより、本実施形態に係る機械学習装置は、処理負荷を低減しつつ、不公平性改善度合いと精度改善度合いとのトレードオフを最適化するようなデータを対象データとして選択することができる。

[0057] さらに、本実施形態に係る機械学習装置は、ラベル無しデータの一部を検証データ、残りを候補データとし、検証データの予測結果と保護属性の値と

の相互情報量の、候補データを与える前後での差分を不公平性改善度合いとして算出する。このように、検証データへの影響を考慮して候補データの不公平性改善度合いが算出されることで、候補データの中から代表性の高いデータが選択され易くなる。

[0058] 図８に、機械学習モデルの予測精度、公平性、及び公平能動学習の実行時間について、本実施形態の手法（以下、「本手法」という）と比較手法との比較を概略的に示す。ここでの比較手法は、非特許文献１に記載の手法のように、ラベル付けの対象となるデータを選択する際に、機械学習モデルの再学習が必要な手法である。本手法は、予測精度及び公平性のトレードオフは比較手法と同等である。また、本手法は、比較手法に比べ、実行時間が大幅に削減される。なお、公平能動学習の理論的な計算コストは、比較手法が「$O((T+N_v)CN_u)$」、本手法が「$O(N_u N_v C^2 M)$」である。なお、Ｔは機械学習モデルの訓練の計算コスト、N_vは検証データの数、N_uは候補データの数、Ｃはラベルの数、Ｍはモンテカルロサンプリング回数である。

[0059] なお、上記実施形態では、ラベル無しデータの一部を検証データとして用いて公平性を評価する場合について説明したが、これに限定されない。ラベル無しデータ集合に含まれる全てのデータを候補データとしてもよい。この場合、公平性の指標としては、例えば、（１）式右辺のΣ内の第２項を除いた式を用いるようにすればよい。また、上記実施形態では、公平性と予測精度とのトレードオフを考慮した評価値を用いてデータを選択する場合について説明したが、公平性のみを考慮してデータを選択するようにしてもよい。

[0060] また、上記実施形態では、ラベル無しデータについての機械学習モデルの予測結果と保護属性の値との独立性を示す指標として相互情報量を用いる場合について説明した。相互情報量は、予測結果と保護属性の値との独立性を数式的に定義できればよく、例えば、カルバックライブラー情報量、イエンセン・シャノン情報量、共分散、Demographic parity difference、Disparate impact ratio等を適用してよい。

[0061] また、上記実施形態では、機械学習プログラムが記憶装置に予め記憶（イ

ンストール）されているが、これに限定されない。開示の技術に係るプログラムは、ＣＤ－ＲＯＭ、ＤＶＤ－ＲＯＭ、ＵＳＢメモリ等の記憶媒体に記憶された形態で提供されてもよい。

符号の説明

[0062] １０　機械学習装置

１１　制御部

１２　訓練部

１４　算出部

１６　選択部

１８　取得部

２０　ラベル付きデータ集合

２２　ラベル無しデータ集合

２４　機械学習モデル

４０　コンピュータ

４１　ＣＰＵ

４２　ＧＰＵ

４３　メモリ

４４　記憶装置

４５　入出力装置

４６　Ｒ／Ｗ装置

４７　通信Ｉ／Ｆ

４８　バス

４９　記憶媒体

５０　機械学習プログラム

５２　訓練プロセス制御命令

５４　算出プロセス制御命令

５６　選択プロセス制御命令

５８　取得プロセス制御命令

６０　情報記憶領域

請求の範囲

[請求項1] ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、前記複数のデータのそれぞれの第１の属性の値との独立性を算出し、

前記独立性に基づいて、前記複数のデータから第１のデータを選択し、

前記第１のデータのラベルを取得し、

前記第１のデータと前記ラベルとに基づいて前記機械学習モデルの訓練を実行する、

処理をコンピュータに実行させることを特徴とする機械学習プログラム。

[請求項2] 前記独立性は、前記予測結果と前記第１の属性の値との相互情報量である請求項１に記載の機械学習プログラム。

[請求項3] 前記第１のデータを選択する処理は、前記複数のデータのそれぞれを前記機械学習モデルに入力した場合の予測結果の不確実性に基づく、前記複数のデータのそれぞれによる前記機械学習モデルの予測精度の改善度合いと、前記独立性とに基づいて選択することを含む、

ことを特徴とする請求項１又は請求項２に記載の機械学習プログラム。

[請求項4] 前記改善度合いは、前記第１のデータが前記機械学習モデルの決定境界に近いほど高くなる、

ことを特徴とする請求項３に記載の機械学習プログラム。

[請求項5] 前記第１のデータを選択する処理は、前記改善度合いと、前記独立性と、前記改善度合いと前記独立性とのトレードオフを表す係数とで表される指標が所定値以上、又は、前記指標が上位所定個のデータを選択することを含む、

ことを特徴とする請求項３に記載の機械学習プログラム。

[請求項6] 前記独立性を算出する処理は、前記複数のデータの一部を検証デー

タ、前記複数のデータのうち前記検証データ以外を候補データとし、前記検証データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記検証データのそれぞれの前記第１の属性の値との独立性であって、前記候補データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記候補データのそれぞれの前記第１の属性の値との独立性に条件付けられた独立性を算出することを含み、

前記第１のデータを選択する処理は、前記候補データから前記第１のデータを選択することを含む、

ことを特徴とする請求項１又は請求項２に記載の機械学習プログラム。

[請求項7] ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、前記複数のデータのそれぞれの第１の属性の値との独立性を算出し、

前記独立性に基づいて、前記複数のデータから第１のデータを選択し、

前記第１のデータのラベルを取得し、

前記第１のデータと前記ラベルとに基づいて前記機械学習モデルの訓練を実行する、

処理をコンピュータが実行することを特徴とする機械学習方法。

[請求項8] 前記独立性は、前記予測結果と前記第１の属性の値との相互情報量である請求項７に記載の機械学習方法。

[請求項9] 前記第１のデータを選択する処理は、前記複数のデータのそれぞれを前記機械学習モデルに入力した場合の予測結果の不確実性に基づく、前記複数のデータのそれぞれによる前記機械学習モデルの予測精度の改善度合いと、前記独立性とに基づいて選択することを含む、

ことを特徴とする請求項７又は請求項８に記載の機械学習方法。

[請求項10] 前記改善度合いは、前記第１のデータが前記機械学習モデルの決定境界に近いほど高くなる、

ことを特徴とする請求項９に記載の機械学習方法。

[請求項11] 前記第１のデータを選択する処理は、前記改善度合いと、前記独立性と、前記改善度合いと前記独立性とのトレードオフを表す係数とで表される指標が所定値以上、又は、前記指標が上位所定個のデータを選択することを含む、

ことを特徴とする請求項９に記載の機械学習方法。

[請求項12] 前記独立性を算出する処理は、前記複数のデータの一部を検証データ、前記複数のデータのうち前記検証データ以外を候補データとし、前記検証データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記検証データのそれぞれの前記第１の属性の値との独立性であって、前記候補データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記候補データのそれぞれの前記第１の属性の値との独立性に条件付けられた独立性を算出することを含み、

前記第１のデータを選択する処理は、前記候補データから前記第１のデータを選択することを含む、

ことを特徴とする請求項７又は請求項８に記載の機械学習方法。

[請求項13] ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、前記複数のデータのそれぞれの第１の属性の値との独立性を算出し、

前記独立性に基づいて、前記複数のデータから第１のデータを選択し、

前記第１のデータのラベルを取得し、

前記第１のデータと前記ラベルとに基づいて前記機械学習モデルの訓練を実行する、

処理を実行する制御部を含むことを特徴とする機械学習装置。

[請求項14] 前記独立性は、前記予測結果と前記第１の属性の値との相互情報量である請求項１３に記載の機械学習装置。

[請求項15] 前記第１のデータを選択する処理は、前記複数のデータのそれぞれ

を前記機械学習モデルに入力した場合の予測結果の不確実性に基づく、前記複数のデータのそれぞれによる前記機械学習モデルの予測精度の改善度合いと、前記独立性とに基づいて選択することを含む、

ことを特徴とする請求項１３又は請求項１４に記載の機械学習装置。

[請求項16] 前記改善度合いは、前記第１のデータが前記機械学習モデルの決定境界に近いほど高くなる、

ことを特徴とする請求項１５に記載の機械学習装置。

[請求項17] 前記第１のデータを選択する処理は、前記改善度合いと、前記独立性と、前記改善度合いと前記独立性とのトレードオフを表す係数とで表される指標が所定値以上、又は、前記指標が上位所定個のデータを選択することを含む、

ことを特徴とする請求項１５に記載の機械学習装置。

[請求項18] 前記独立性を算出する処理は、前記複数のデータの一部を検証データ、前記複数のデータのうち前記検証データ以外を候補データとし、前記検証データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記検証データのそれぞれの前記第１の属性の値との独立性であって、前記候補データのそれぞれを前記機械学習モデルに入力した場合の予測結果と、前記候補データのそれぞれの前記第１の属性の値との独立性に条件付けられた独立性を算出することを含み、

前記第１のデータを選択する処理は、前記候補データから前記第１のデータを選択することを含む、

ことを特徴とする請求項１３又は請求項１４に記載の機械学習装置。

[請求項19] ラベル付けされていない複数のデータのそれぞれを機械学習モデルに入力した場合の予測結果と、前記複数のデータのそれぞれの第１の属性の値との独立性を算出し、

前記独立性に基づいて、前記複数のデータから第１のデータを選択

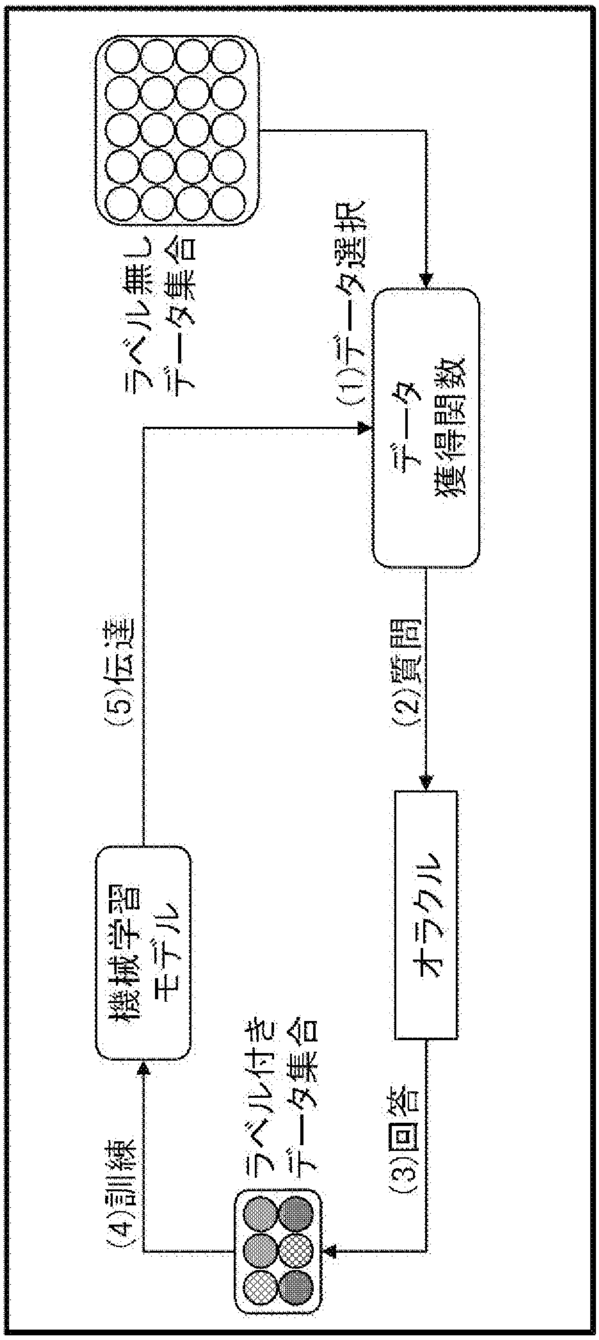
し、

前記第１のデータのラベルを取得し、

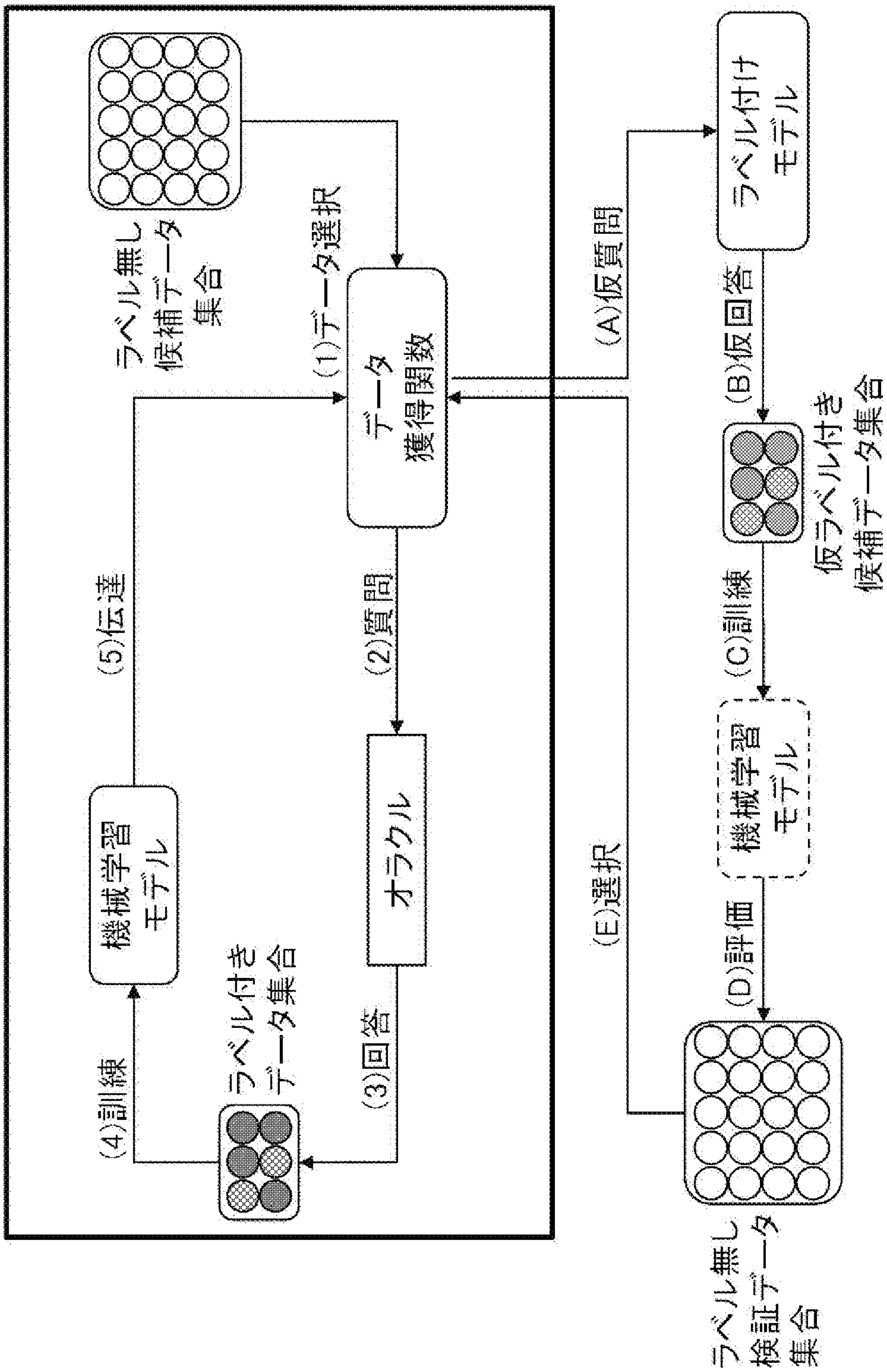
前記第１のデータと前記ラベルとに基づいて前記機械学習モデルの訓練を実行する、

処理をコンピュータに実行させることを特徴とする機械学習プログラムを記憶した非一時的記憶媒体。

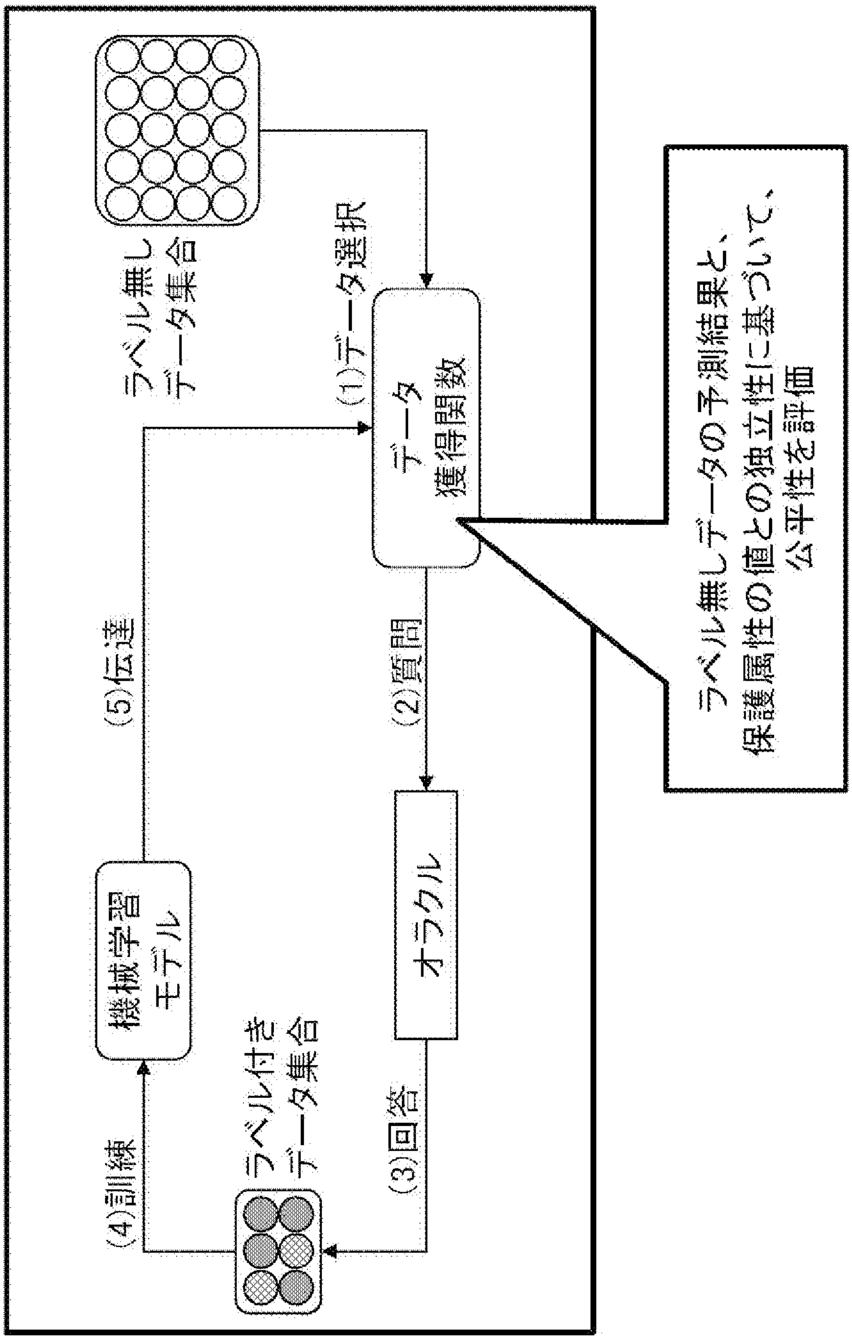
[図1]


ラベル無し
データ集合
(1)データ選択
データ
獲得関数
(2)質問
オラクル
(3)回答
ラベル付き
データ集合
(4)訓練
機械学習
モデル
(5)伝達

[図2]

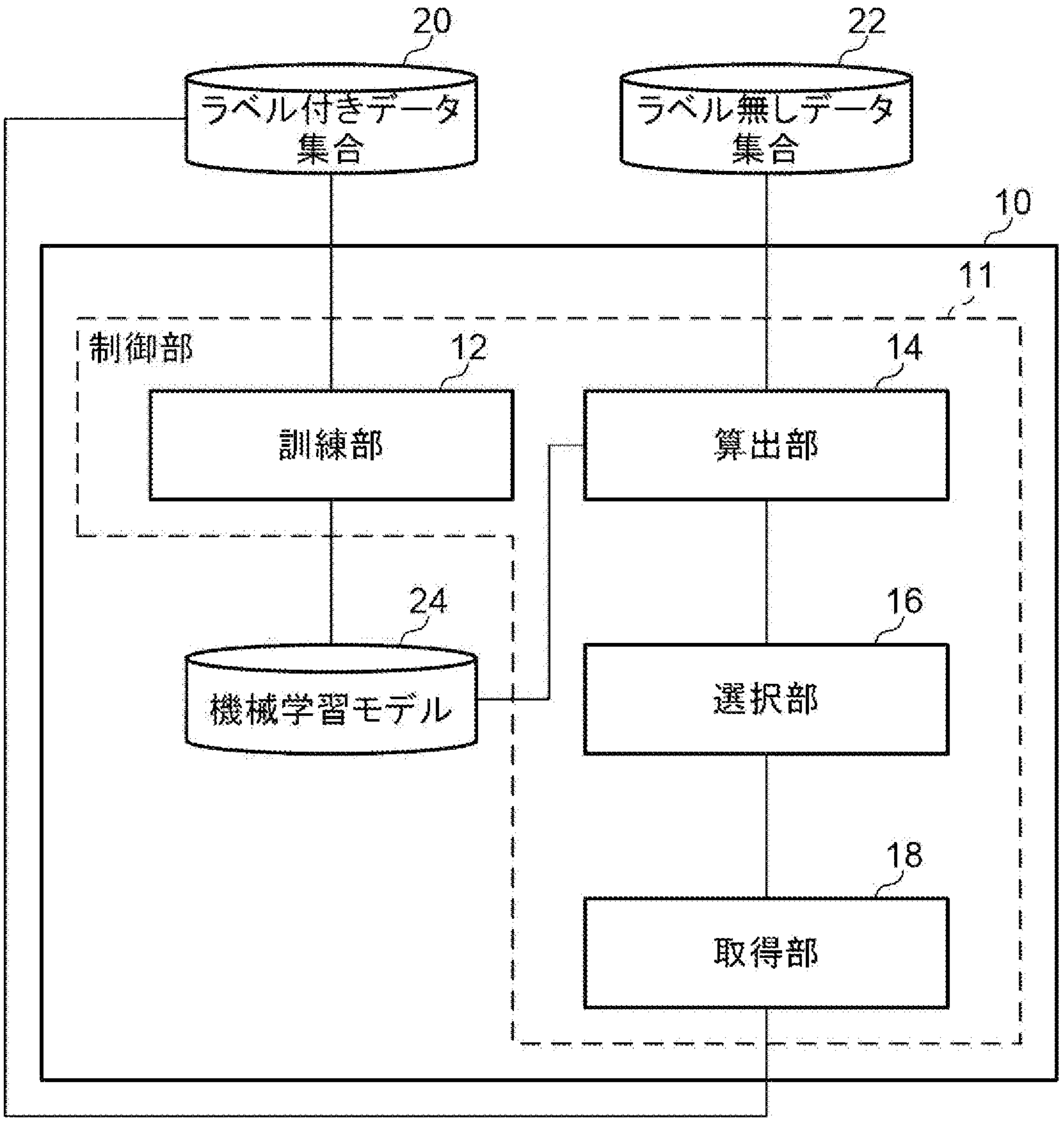

ラベル無し候補データ集合
(1)データ選択
データ獲得関数
(2)質問
オラクル
(3)回答
ラベル付きデータ集合
(4)訓練
機械学習モデル
(5)伝達
(A)仮質問
ラベル付けモデル
(B)仮回答
仮ラベル付き候補データ集合
(C)訓練
機械学習モデル
(D)評価
ラベル無し検証データ集合
(E)選択

[図3]

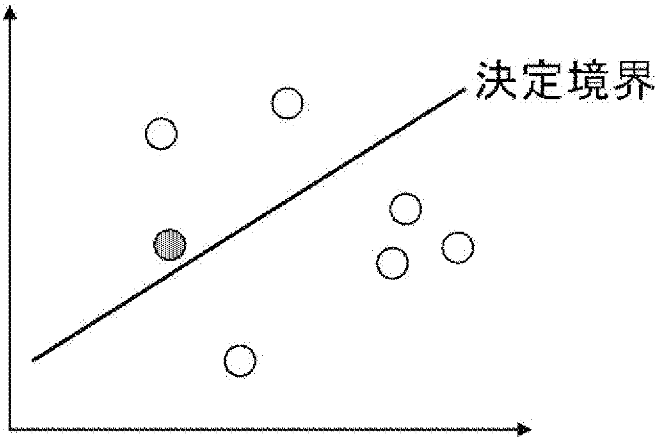

ラベル無し
データ集合
(1)データ選択
データ
獲得関数
(2)質問
オラクル
(3)回答
ラベル付き
データ集合
(4)訓練
機械学習
モデル
(5)伝達
ラベル無しデータの予測結果と、
保護属性の値との独立性に基づいて、
公平性を評価

WO 2024/166331 PCT/JP2023/004458

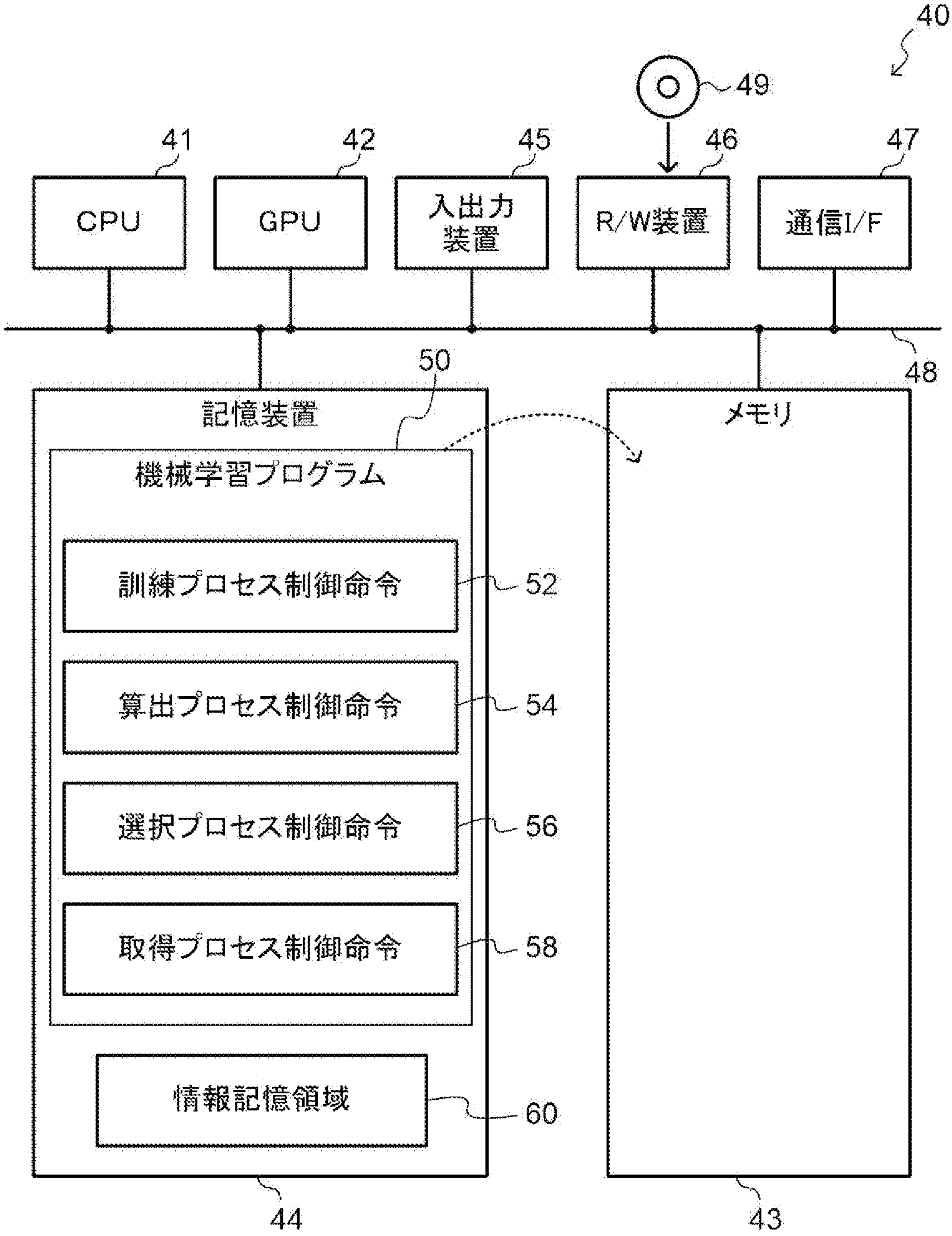
[図4]



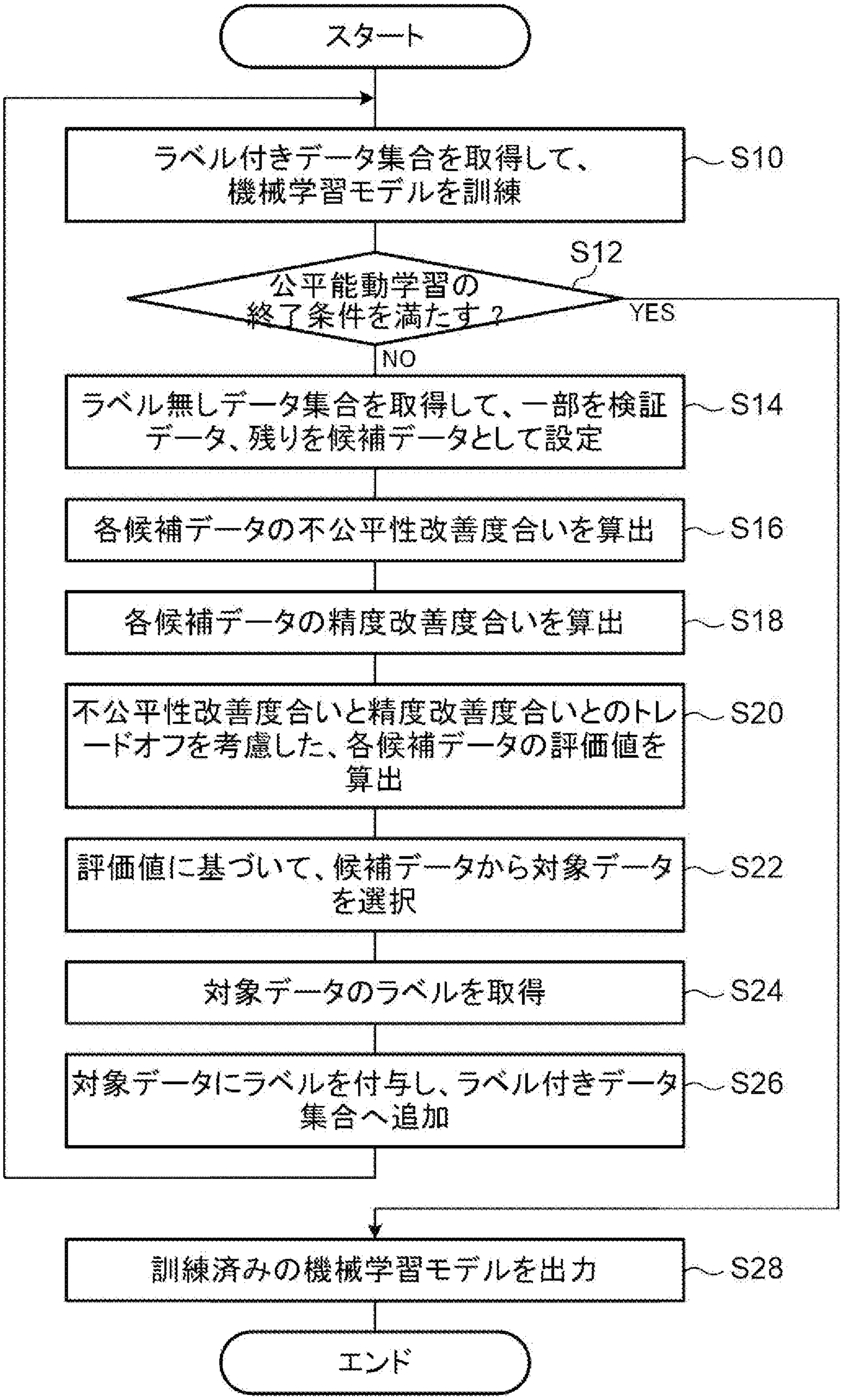
[図5]



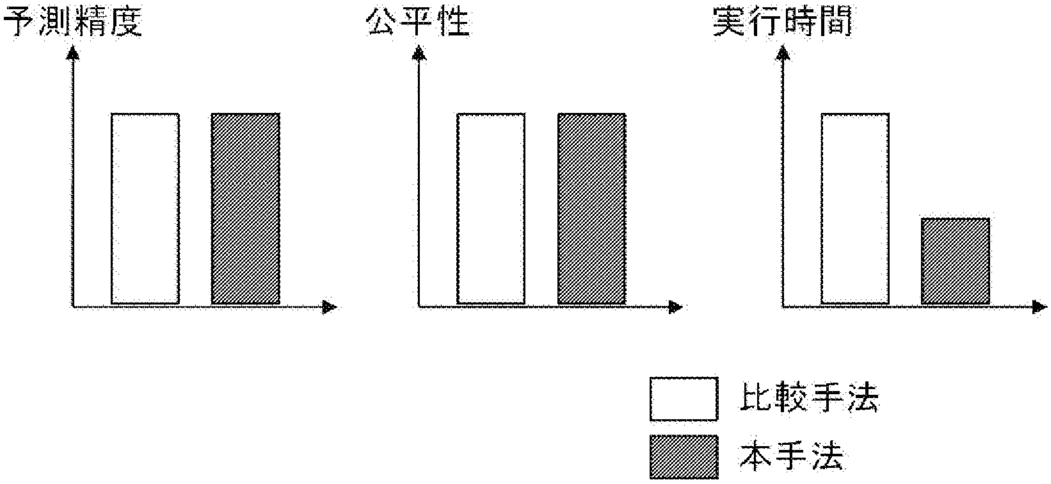
[図6]


40
49
41
42
45
46
47
CPU
GPU
入出力
装置
R/W装置
通信I/F
48
50
記憶装置
メモリ
機械学習プログラム
訓練プロセス制御命令
52
算出プロセス制御命令
54
選択プロセス制御命令
56
取得プロセス制御命令
58
情報記憶領域
60
44
43

［図7］


スタート
ラベル付きデータ集合を取得して、
機械学習モデルを訓練
S10
S12
公平能動学習の
終了条件を満たす？
YES
NO
ラベル無しデータ集合を取得して、一部を検証
データ、残りを候補データとして設定
S14
各候補データの不公平性改善度合いを算出
S16
各候補データの精度改善度合いを算出
S18
不公平性改善度合いと精度改善度合いとのトレ
ードオフを考慮した、各候補データの評価値を
算出
S20
評価値に基づいて、候補データから対象データ
を選択
S22
対象データのラベルを取得
S24
対象データにラベルを付与し、ラベル付きデータ
集合へ追加
S26
訓練済みの機械学習モデルを出力
S28
エンド

[図8]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2023/004458

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/091(2023.01)i; ***G06N 20/00***(2019.01)i
FI: G06N3/091; G06N20/00 130

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)

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 1971-2023
Registered utility model specifications of Japan 1996-2023
Published registered utility model applications of Japan 1994-2023

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	ANAHIDEH, Hadis et al., Fair Active Learning, arXiv [online], 31 March 2021, [retrieved on 20 April 2023], Internet: <URL: https://arxiv.org/abs/2001.01796v5> entire text, all drawings	1-19
A	US 2021/0035014 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 04 February 2021 (2021-02-04) entire text, all drawings	1-19
A	WO 2022/254626 A1 (FUJITSU LIMITED) 08 December 2022 (2022-12-08) entire text, all drawings	1-19

☐ 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
20 April 2023	**09 May 2023**

Name and mailing address of the ISA/JP	Authorized officer
Japan Patent Office (ISA/JP) **3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915** **Japan**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/JP2023/004458

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2021/0035014 A1	04 February 2021	(Family: none)	
WO 2022/254626 A1	08 December 2022	(Family: none)	

国際調査報告

国際出願番号 PCT/JP2023/004458

A. 発明の属する分野の分類（国際特許分類（ＩＰＣ））
G06N 3/091(2023.01)i; G06N 20/00(2019.01)i FI: G06N3/091; G06N20/00 130

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））	
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	１９７１－２０２３年
日本国実用新案登録公報	１９９６－２０２３年
日本国登録実用新案公報	１９９４－２０２３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	ANAHIDEH, Hadis et al., Fair Active Learning, arXiv [online], 2021.03.31, [検索日 2023.04.20], インターネット：<URL:https://arxiv.org/abs/2001.01796v5> 全文、全図	1-19
A	US 2021/0035014 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 04.02.2021 (2021－02－04) 全文、全図	1-19
A	WO 2022/254626 A1 (富士通株式会社) 08.12.2022 (2022－12－08) 全文、全図	1-19

☐ 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

“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
20.04.2023	09.05.2023
名称及びあて先	権限のある職員（特許庁審査官）
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	山本　俊介 5B 5087 電話番号 03-3581-1101 内線 3545

国際調査報告
パテントファミリーに関する情報

国際出願番号
PCT/JP2023/004458

引用文献	公表日	パテントファミリー文献	公表日
US 2021/0035014 A1	04.02.2021	(ファミリーなし)	
WO 2022/254626 A1	08.12.2022	(ファミリーなし)	